Protein-Protein Interaction Network Alignment and Evolution

by

Brian Man-Kin Law

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Computer Science
University of Toronto

Protein-Protein Interaction Network Alignment and Evolution

Brian Law

Doctor of Philosophy

Computer Science
University of Toronto

2019

Abstract

Network alignment is an emerging analysis method enabled by the rapid large-scale collection of protein-protein interaction data for many different species. As sequence alignment did for gene evolution, network alignment will hopefully provide new insights into network evolution and serve as a new bioinformatic tool for making biological inferences across species.

Using new SH3 binding data from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*, I construct new interface-interaction networks and devise a new network alignment method for these networks. With appropriate parameterization, this method is highly successful at generating alignments that reflect known protein orthology information and contain high network topology overlap. However, close examination of the optimal parameterization reveals a heavy reliance on protein sequence similarity and fungibility of other data features, including network topology data, an observation that may also pertain to protein-protein interaction network alignment.

Closer examination of interactomic data, along with established orthology data, reveals that protein-protein interaction conservation is quite low across multiple species, suggesting that the high network topology overlap achieved by contemporary network aligners is ill-advised if biological relevance of results is desired. Further consideration of gene duplication and protein

binding sites reveal additional PPI evolution phenomena further reducing the network topology overlap expected in network alignments, casting doubt on the utility of network alignment metrics solely based on network topology.

Instead, I suggest a new framework to think about protein-protein interaction network alignment focused on generating and validating small-scale inferences. I create a prototype alignment visualization and analysis tool to facilitate this approach, which will hopefully aid researchers in learning more about the mechanisms of network evolution and how network alignment can model them.

Acknowledgments

I would like to thank my supervisor Gary Bader, without whom this thesis would never exist. His unfailing confidence and patience were my constant companions on this long journey, whether I was working myself into a frenzy or paralyzed by doubt. Even when I felt it unwarranted, Gary's belief in me never wavered, and without his encouragement, I would have turned off this path long ago. Thank you.

I would like to thank the members of my committee Derek Corneil, Alan Moses, and Zhaolei Zhang for their insight and support. Their questions and ideas have occupied me for many years, and I regret I still cannot provide the answers that they deserve. The challenges they have laid out for me have driven, inspired, and haunted me over the years, and I will continue striving to meet them. Thank you.

I would like to thank the graduate office staff at the Department of Computer Science, past and present: Julie Weedmark, Linda Chow, Kolden Simmons, Vinita Krishnan, Margaret Meaney, Celeste Francis Esteves, Erin Bedard, Vivian Hwang, and Lynda Barnes, for their assistance in navigating the rocky shoals of academic bureaucracy. I am unsure any graduate student could finish their degrees in our department without their assistance, and yet their response to my incompetence with paperwork was never anything but helpfulness. Thank you.

I would like to thank the many people, too many people to name, who have touched my life positively and kept me on track these past years. My fellow Computer Science graduate students, my CCBR 6th floor-mates, my Bader Lab compatriots, the many undergraduate students I have had the pleasure of teaching, my CUPE 3902 comrades, my teaching faculty mentors, my friends and my family. This has been a long and difficult labour, and it may have entirely consumed me but you reminded me that there was more to do, more to me than simply my work. Thank you.

Finally, I would like to apologize to my grandmother, (Monica) Tack Ching Soong. I thought I would have more time to make you proud of me. I took it for granted while you raised me, cared for me, nurtured me. I never expected it would dwindle so quickly, and now that I am out of time, I find it supplanted by regret, so many regrets, big and small. I am sorry.

Table of Contents

Acknowledgments iv

Table of Contents v

List of Tables ix

List of Figures x

List of Appendices xiii

List of Abbreviations xiv

Chapter 1 1

1 Background 2

1.1 Introduction 2

1.1.1 Comparative Evolution 2

1.1.2 Protein-Protein Interactions 3

1.1.3 Protein-Protein Interaction Data Quality 5

1.1.4 Interface-Interaction Networks 6

1.2 Protein-Protein Interaction Network Alignment 11

1.2.1 Theoretical Considerations 12

1.2.2 Local Network Alignment 13

1.2.3 Global Network Alignment 14

1.2.4 Assessment and Evaluation of Network Alignments 17

1.3 Protein-Protein Interaction Network Evolution 27

1.3.1 Fundamentals and Applications 27

1.3.2 Models for Protein-Protein Interaction Networks 29

1.4 Protein-Protein Interaction Network Alignment Visualization 31

Chapter 2 33

2 Alignment of Interface-Interaction Networks 34

2.1 Introduction 34

2.1.1 Network Alignment Theory 34
2.1.2 Interface-Interaction Networks 35
2.1.3 Protein-Protein Interaction Network Alignment 36
2.2 Results 39
2.2.1 Comparison with PPIN Alignment Algorithms 39
2.2.2 Incorporating More Similarity Features 45
2.2.3 Parameter Weight Tuning 47
2.2.4 A Zoom-in 48
2.2.5 Yeast Subspecies Alignments 51
2.3 Discussion 53
2.4 Conclusions 58
2.5 Methods 58
2.5.1 Algorithm 58
2.5.2 Comparison Algorithms 61
2.5.3 Network Creation 63
2.5.4 Orthology Data 64
2.5.5 Similarity Feature Data 64
2.5.6 Parameter Training Procedure 65
2.5.7 Similarity Feature Reduction 65
Chapter 3 67
3 Dynamics of Protein-Protein Interaction Conservation 68
3.1 Introduction 68
3.2 Results 70
3.2.1 Interolog conservation across species 70
3.2.2 Impact of PRMs on Protein Interactivity 81

3.2.3 The importance of domain architecture in protein-protein interaction conservation 87
3.3 Discussion 95
3.3.1 Local is better than global for network alignment 96
3.3.2 PPI Data Quality 97
3.3.3 Quality of binding site data 98
3.4 Conclusion 98
3.5 Methods 99
3.5.1 Protein-protein Interaction Data 99
3.5.2 Domain Data 100
3.5.3 SH3 Domain Binding Data 101
Chapter 4 102
4 Small-Scale Visualization of Protein-Protein Interaction Network Alignment 103
4.1 Introduction 103
4.2 Results 104
4.2.1 Description of PPAAT 104
4.2.2 Use Cases 107
4.3 Discussion 112
4.3.1 Additional Features to be Implemented 112
4.3.2 The Broader PPAAT Framework: Rethinking Network Alignment 113
4.4 Conclusion 114
Chapter 5 115
5 Summary and Future Directions 116
5.1 Summary of Thesis 116
5.2 Future Directions 117
5.2.1 Full Development and Deployment of PPAAT 117

5.2.2 Interaction/Edge Attributes 118
5.2.3 Critical Assessment and Validation of Network Alignment Methods 119
5.2.4 Modelling the Mechanisms of Network Evolution 120
References 122
Copyright Acknowledgements 136
Appendices 137

List of Tables

Table 2-1 – Comparison between GreedyPlus, C-GRAAL, IsoRank, and Natalie 2.0 with *C. elegans* and *S. cerevisiae* SH3-mediated IINs. 41

Table 2-2 – Comparison between GreedyPlus, C-GRAAL, GRAAL, and H-GRAAL with *C. elegans* and *S. cerevisiae* SH3-mediated IINs. 44

Table 2-3 – Alignment algorithm performance on *C. elegans* and *S. cerevisiae* SH3-mediated IINs using all similarity features. 46

Table 2-4 – An "optimal" parameter set for GreedyPlus, normalized out of 100. 47

Table 2-5 – A reduced "optimal" parameter set for GreedyPlus, normalized out of 100. 48

Table 2-6 – GreedyPlus performance using different similarity features. 57

Table 3-1 – Correlation coefficients between the number of interaction partners for all pairs of protein orthologs in various pairs of species. 72

Table 3-2 – Interaction conservation between five model organism species (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *S. cerevisiae*). 77

Table 3-3 – Correlation coefficients between the number of domains in a protein and the number of PPIs involving that protein, aggregated per species, using different domain types. 82

Table 3-4 – Evolutionary gain/loss events of SH3 domains. 88

Table 3-5 – Number of interologs shared between intersectin orthologs. 91

Table 3-6 – Multiple sequence alignments of human ITSN1/2 SH3 binding site regions with yeast and worm orthologs. 94

Table 3-7 – Summary of interaction data used. 99

Table 3-8 – Summary of SH3 complements for several species. 101

List of Figures

Figure 1-1 – An example of the additional detail provided by interface-interaction networks. 8

Figure 1-2 – Figure 1 from *Comprehensive Analysis of the Human SH3 Domain Family Reveals a Wide Variety of Non-canonical Specificities*. 10

Figure 1-3 – A representative subnetwork of the S. cerevisiae SH3-mediated interface interaction network. 11

Figure 1-4 – Supplementary Figure 3 from *Fuse: multiple network alignment via data fusion*, including original caption. 19

Figure 1-5 – Supplementary Figure S9 from *MAGNA: Maximizing Accuracy in Global Network Alignment*, including original caption. 20

Figure 1-6 – Supplementary Figure S13 from *MAGNA: Maximizing Accuracy in Global Network Alignment*, including original caption. 21

Figure 1-7 – Figure 4 from *Unified Alignment of Protein-Protein Interaction Networks*, including original caption. 22

Figure 1-8 – Table 1 from *A comparison of algorithms for the pairwise alignment of biological networks*, including original caption. 23

Figure 1-9 – Part of Supplementary Figure 2 from *Unified Alignment of Protein-Protein Interaction Networks*. 26

Figure 1-10 – Figure 1 from *GASOLINE: a Cytoscape app for multiple local alignment of PPI networks*. 32

Figure 2-1 – Illustrative examples of represented protein orthologies (RPOs) and orthologous node pairs (ONPs). 40

Figure 2-2 – IsoRank alignment of worm and yeast SH3-mediated IINs, using only protein BLAST. 42

Figure 2-3 – GreedyPlus alignment of worm and yeast SH3-mediated IINs, using only protein BLAST, EAW = 0.5. 44

Figure 2-4 – A density plot of graphlet similarity scores between orthologous nodes and random node pairs. 45

Figure 2-5 – A zoom-in of the "optimal" GreedyPlus alignment of worm and yeast SH3-mediated IINs, consisting of the two yeast BZZ1 nodes and all their neighbours. 49

Figure 2-6 – A zoom-in of the "optimal" GreedyPlus alignment of worm and yeast SH3-mediated IINs, consisting of the three worm CYK1 nodes and all their neighbours. 50

Figure 2-7 – Percent RPO and EA achieved for pairwise yeast species alignments. 52

Figure 2-8 – Percent RPO and EA achieved for pairwise yeast species alignments. 53

Figure 2-9 – Trade-off in GreedyPlus performance between orthologous nodes aligned and edges aligned. 55

Figure 2-10 – A simple example of GreedyPlus in action. 60

Figure 2-11 – Pseudocode for the GreedyPlus algorithm. 61

Figure 2-12 – The Greedy algorithm for IIN alignment. 62

Figure 2-13 – The Seed and Extend algorithm for IIN alignment. 63

Figure 3-1 – Comparisons of the number of interaction partners for all pairs of protein orthologs in various pairs of species. 71

Figure 3-2 – Illustration of various interolog scenarios. 73

Figure 3-3 – Boxplots and density plots comparing the mean BLASTP bit-score between interologous and non-interologous protein pairs between selected pairs of species. 75

Figure 3-4 – The rate of conservation for PPIs from different species in the *H. sapiens* interactome. 79

Figure 3-5 – % interolog conservation for PPIs in the *H. sapiens* interactome, grouped by the number of potential interologs. 80

Figure 3-6 – Scatterplots relating the number of SH3 domains and the number of PPIs for orthologous protein pairs across different species. 83

Figure 3-7 – Scatterplots relating the number of PRMs and the number of PPIs for orthologous protein pairs across different species. 85

Figure 3-8 – Schematic of interaction conservation between yeast EDE1 and worm ITSN-1..... 93

Figure 3-9 – Schematic of interaction conservation between human ITSN1 and worm ITSN-1. 95

Figure 4-1 – The default PPAAT view. 105

Figure 4-2 – A reduced PPAAT view. 106

Figure 4-3 – Dropdown menu for proteins in PPAAT. 107

Figure 4-4 – A PPAAT visualization of paralogs *C. elegans* ITSN-1 and *H. sapiens* ITSN2. .. 108

Figure 4-5 – A PPAAT visualization of non-homologous proteins *S. cerevisiae* BOI1 and *C. elegans* Y44E3A.4. 109

Figure 4-6 – PPAAT Predicted PPIs for *C. elegans* ITSN-1 and *H. sapiens* ITSN1 based on interolog conservation. 110

List of Appendices

A Additional Figures Demonstrating the Impact of Gene Duplication on Interolog Conservation 138

B Protein-Protein Interactions of Intersectin and Orthologs 146

List of Abbreviations

BLAST: Basic Local Alignment Search Tool

DD: duplication-divergence

EA: edges aligned

EAW: edge alignment weight

EC: edge coverage

FPR: false positive rate

GNA: global network alignment

GO: Gene Ontology

IIN: interface-interaction network

KEGG: Kyoto Encyclopedia of Genes and Genomes

KP: Kyoto Encyclopedia of Genes and Genomes pathway

LNA: local network alignment

LOWESS: locally weighted scatterplot smoothing

MS: mass spectrometry

ONP: orthologous node pairs

PCA: protein complementation assay

PPAAT: Pairwise Protein Alignment Analysis Tool

PPI: protein-protein interaction

PPIN: protein-protein interaction network

PRM: peptide recognition module

RPO: represented protein orthology

TCSS: Topological Clustering Semantic Similarity

Y2H: yeast two-hybrid

Chapter 1

Background

1 Background

1.1 Introduction

1.1.1 Comparative Evolution

Evolution is the fundamental organizing principle in modern biology. The markers of conservation and divergence are some of the few sensible patterns by which biological knowledge can be organized. Where conservation is found, we can infer the existence of selective forces acting to preserve some key functionality. Where divergence is found, we can hypothesize the rise of some new development, either intrinsic or extrinsic. Observation of both conservation and divergence requires comparison, which occupies an exceptionally prominent role in biological data analysis.

Alignment is a uniquely biological approach to comparative analysis. The sequence alignment algorithms Needleman-Wunsch[1] and Smith-Waterman,[2] the sequence alignment search heuristic BLAST[3], and multiple sequence alignment algorithms like MUSCLE[4] are ubiquitous and indispensable tools in computational biology. Aligning genes/proteins with these tools has been essential for understanding their function, such as by identifying protein domains, the structural, functional and evolutionary units that make up a large part of proteins. Structural alignment of proteins has also proven a similarly important tool.[5]

Sequence and structure data are, however, alone inadequate for study of complex biological systems, because they do not naturally capture the interactions between genes/proteins. Scientists interested in understanding complex biological systems have had to create new methods and models to represent these systems. Commonly, this is done by considering several proteins that collaborate for a specific cellular process, known as a *functional module*. These modules include complexes, wherein many proteins gather together to perform a particular function, and pathways, wherein proteins are organized in a process to perform a particular function.

Detection and confirmation of protein complexes and pathways, however, was difficult using traditional experimental methods. Protein complexes needed to be purified and crystallized, a difficult and costly procedure that is highly sensitive to chemical conditions. Pathway detection often involved multiple experiments, such as using gene coexpression experiments to identify putative pathway participants and then performing gene knockout experiments to confirm their

functional cooperation. These requirements substantially limited the number of pathways and complexes that could be identified and characterized, which then further limited comparative study of these modules between species.

Now, with the arrival of high-throughput protein-protein interaction (PPI) methods, we have entered a new period of big data comparative systems biology. Instead of experimental data generation lagging badly behind hypothesis generation, PPI data is now generated en masse, awaiting analysis and exploitation. While the data is still far from perfect or complete, its plentifulness has ushered in a new approach to systems biology, one which is expansive and ambitious, limited more by what can be understood rather than what is prohibitively costly.

1.1.2 Protein-Protein Interactions

Proteins *in vivo* need to cooperate in order to perform the myriad biological functions required to sustain life. This cooperation often occurs via physical interaction between protein molecules, called protein-protein interactions (PPIs). These interactions can take a variety of forms, ranging from transient interactions to longer-term stable interactions, between pairs of proteins or as multi-protein complexes, and can include self-interactions such as in homodimers. Scientists have long organized proteins into discrete functional units, known as modules, such as pathways and complexes for study. These modules are bound together by PPIs, and in this manner, PPIs have implicitly served as a foundational element of systems biology, but they were not considered independently and broadly.

Over time though, PPI data has been gathered in ever-increasing quantities, due to the development of high-throughput methods for the detection of PPIs, or PPI mapping, such as yeast-two-hybrid (Y2H), protein complementation assays (PCA), and affinity purification followed by mass spectrometry (AP-MS). There has been dramatic growth recently, in both the number of species with PPI data and the number of PPIs for each species, as PPI mapping experiments have proliferated.

With this newfound availability of PPI data, scientists have begun to focus on PPIs collectively, to see what biological knowledge can be learned from direct study of PPIs and incorporation of PPI data with other biological data. The set of all PPIs in a given species is known as the *interactome*, which can be represented as a protein-protein interaction network (PPIN), wherein

the nodes are proteins and the edges interactions between physically interacting proteins. PPINs are the only way currently available to model most, if not all, the biological systems within the entire cell simultaneously. While there are significant limitations in relying solely on PPI data for such modelling, there is immense power in collecting an entire interactome into a single network for analysis.

The number of computational biology tools for PPIs and PPINs has increased greatly in recent years. These include visualization software like Cytoscape,[6] PPI predictors like PIPE,[7,8] and PPI databases such as BioGRID, iRefIndex, and others.[9-20] PPI data can also be used to infer protein essentiality[21] and identify disease-causing genes and mutations.[22-25]

Perhaps the most popular usage for PPI data is for gene function prediction. Traditionally, gene function predictions were made based on the sequence homology between a protein of known function and a protein of unknown function. Conservation of sequence implied conservation of function. However, this approach is relatively limited, due to the limited number of homologous protein pairs matching this specific profile. With PPI data, because proteins often interact with one another in service of a biological function, an uncharacterized protein's function can often be inferred via its interactions with characterized proteins with known function.[26,27] This is known as the "guilt by association" model, and is used by protein function predictors such as GeneMANIA, which integrates PPI data with gene expression and other data to predict protein functions across the entire proteome.[28]

Despite the power of and interest in PPI data for driving biological research, it must be noted that PPIs are a crude model of the biological systems working in the cell. There are a number of representational limitations when modeling PPIs as a network, such as the inability of PPI data to capture interactions between proteins and other types of molecules, such as ribosomal RNA, which may hide indirect but important interactions between protein molecules.

There are also abstractions made due to the limits of current PPI mapping technology and the desire to create comprehensive datasets. For example, PPI data is aggregated irrespective of cell type and the cellular conditions under which the data was collected, despite the fact that they can greatly impact gene expression, alternative splicing, and post-translational modification. Furthermore, while PPI mapping technology has seen many recent developments, PPI characterization studies have not kept up, and so very little is known about the nature of many

PPIs, such as whether they are permanent or transient, strong or weak, or dependent on specific cellular conditions. Finally, PPI data is often organized around genes rather than proteins, thus sidestepping the issues of how to select the correct protein isoform, in the common case of multiple proteins encoded by the same gene, and how to deal with less reliable, in comparison to genetic data, proteomic data.

While some of these limitations may yet be overcome with technological advancements, others may be entirely insurmountable. What the consequences of these limitations are is yet unclear; only by continuing to utilize the data will we encounter the limitations of our current approaches to using PPIs to learn about biology, and discover how we might rectify some of the shortcomings of current approaches.

1.1.3 Protein-Protein Interaction Data Quality

Despite major advances in PPI mapping methods, there remain major uncertainties when working with PPI data. Generally, a single, experimental hit in a PPI database is sufficient for a PPI to be considered true, an approach that may be particularly susceptible to false positives. Troublingly, studies indicate that Y2H, PCA, and MS detect highly disparate sets of PPIs, so they are poorly suited for mutual validation.[29] Additionally, even estimating the rate of false positives is difficult, due to how PPI data is aggregated. In the absence of a comprehensive PPI mapping project, existing PPI data comes from a pastiche of uncoordinated PPI mapping projects, each using slightly different protocols. Nor is there an established consensus on the false positive rate (FPR) for experimentally detected PPIs, as early estimates of species interactome sizes have long since been exceeded.[20,30]

Similarly, false negative data likely troubles current PPI datasets. With scientists independently conducting PPI mapping experiments, there is significant social bias in the proteins selected for experimentation, based on their biomedical relevance and ease of availability.[31,32] Proteins not covered by PPI mapping experiments cause inexplicable gaps in the known interactome, a phenomenon that is particularly noticeable when working with smaller interactome datasets, such as when Zhang et al. added 1680 FANCD2 PPIs to BioGRID's *M. musculus* PPIN,[33] which at the time only held 39 146 PPIs,[20,34] for a sudden 4.3% growth, when McFarland et al. added 171 SNCA PPIs in 2008,[35,36] or when Piazzi et al. added 188 PLCB1 PPIs in 2013.[37,38] These sudden

additions could dramatically affect both past and future results that are based on specific database versions, hampering reproducibility, biasing analyses, and prompting false conclusions.

Furthermore, because PPI detection methods are not consistently effective for different PPIs, a single study may be insufficient to fully map a protein's PPIs, which instead would require coverage from multiple experiments with different methods. For example, membrane proteins are a large fraction of the proteome, nearly 30%, but are badly underrepresented in PPI databases because of the difficulty in applying established PPI detection methods to them.[39-41] Determining whether a protein truly has few PPIs or was not covered by PPI mapping experiments is difficult because PPI databases do not store negative results, except small specialized databases such as Negatome.[42] Consequently, the only way to definitively identify a non-interaction from the literature is an in-depth review of PPI mapping studies, a technique at odds with the "big data" approach often embodied by PPI analyses.

Thus, while the availability of PPI data is increasing rapidly, there remain persistent data quality issues that may not be resolved simply by conducting more PPI mapping studies. Generating more PPI data will not innately improve the reliability of the data nor establish a known quality level for the data. Instead, further PPI mapping studies may need to be coordinated to be mutually validating, between both high-throughput and low-throughput methods so as to establish their discrepant capabilities and to distinguish those discrepancies from technical error or experimental incompleteness. By increasing the depth of PPI mapping coverage, rather than simply its width, we can hopefully improve our certainty of PPI data quality. Until then, though PPI data holds great potential for scientific discovery, analyses must be carefully designed to avoid erroneous conclusions resulting from variable data quality.[43]

1.1.4 Interface-Interaction Networks

One key abstraction in PPIN models is that proteins are treated as monadic. In reality, PPIs typically are not mediated by the entirety of the participating proteins; instead, they operate between specific binding interfaces on those proteins. Interface-interaction networks (IINs) are a refinement of PPINs wherein proteins are subdivided into their separate interaction interfaces.[44] IINs can be represented using a traditional graph model where a node represents a specific binding site and an edge represents a physical interaction between two binding interfaces on their respective proteins, or a hypergraph model in which protein nodes contain interface nodes. In the

hypergraph model, IINs are a higher resolution version of the PPIN, layering on additional binding interface information where available.

The higher resolution of IINs allows for new biological insights that cannot be derived from standard PPINs. For example, IINs can distinguish between "date hubs" – proteins that interact with many partners, but at different times or in different locations – and "party hubs" – proteins that interact with many partners simultaneously (see Figure 1-1).[45] While these distinct types of hub proteins will appear identically in a PPIN, in an IIN, the former will have few binding sites that are reused for many different interaction partners whereas the latter will have many binding sites that are specific for each interaction partner. This is useful to help elucidate the evolutionary processes and constraints acting on hub proteins. The study of IINs will also help interpret how protein domain and binding site gain and loss affect the PPIN, predict PPIN perturbations caused by sequence mutations that affect binding sites, and allow in-depth analysis of how protein-protein interactions are formed and lost.[46-48]

Network topology differences between IINs and PPINs, however, mean that algorithms designed to operate on PPINs may not function properly with IINs. While PPINs are often sparse, IINs are much more so, with each PPIN node (protein) split into multiple nodes that represent the different binding sites on that protein. Similarly, while PPINs exhibit a hub and spoke topology, with many low-degree and fewer high-degree nodes, this characteristic is exaggerated in IINs. For example, protein-recognition modules, such as protein kinases or SH3 domains, are often capable of binding many different proteins, leading to relatively few high-degree nodes connected to many low degree nodes. Additionally, due to binding specificity similarities, different domains will often recognize the same ligands, forming a multi-fan network topology. Methods that depend on the neighbourhoods of nodes being topologically distinct to generate their alignments get confused by these repeated patterns and thus perform inconsistently (see below for examples).

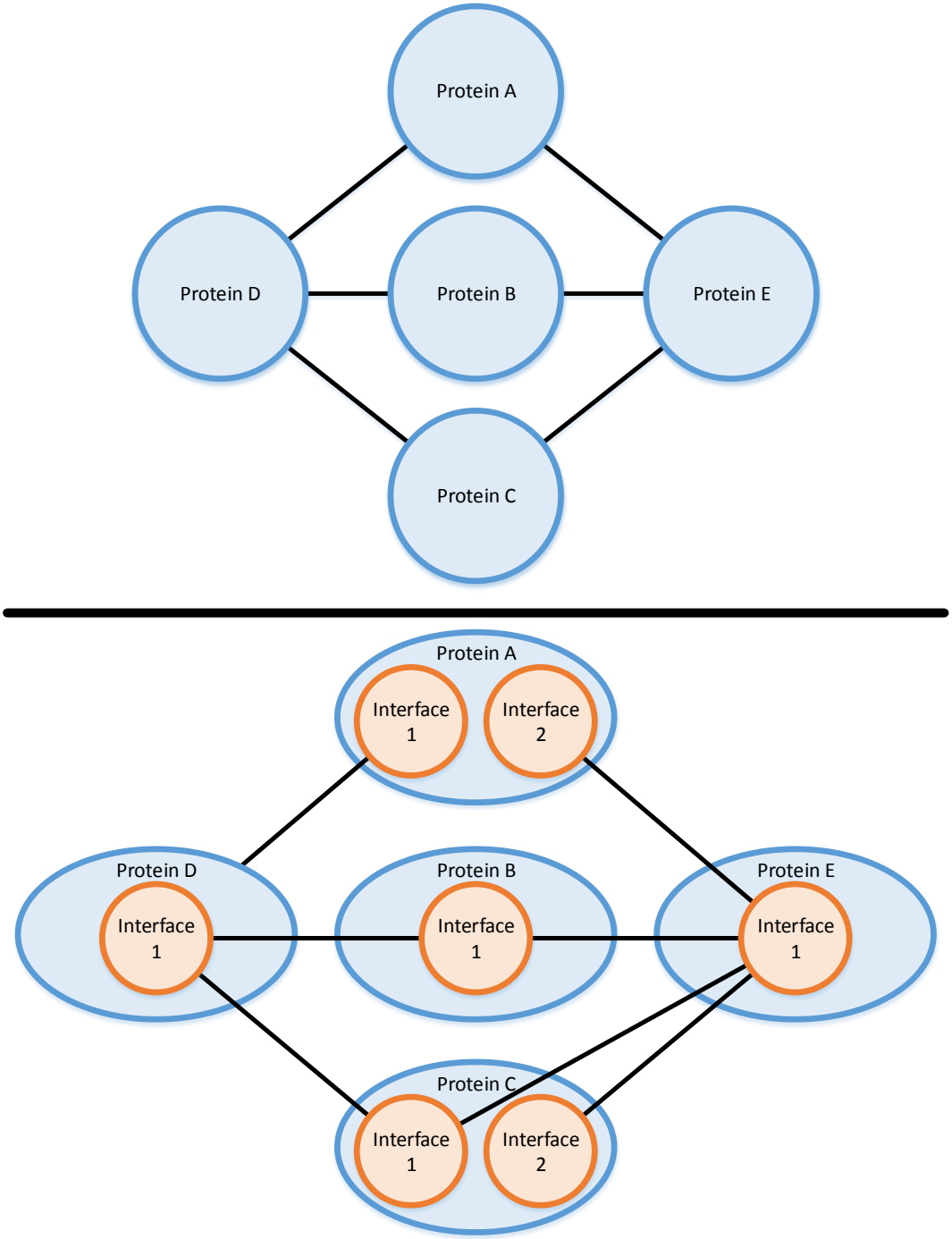


Figure 1-1 – An example of the additional detail provided by interface-interaction networks. In a traditional protein-protein interaction network (top), proteins A, B, and C appear identical, based on their interactions with proteins D and E. However, with additional interface-specific information, as shown in the interface-interaction network (bottom), they can be distinguished. Protein A has two distinct interfaces that accommodate binding with proteins D and E, indicating that concurrent binding may be possible. Protein B has only a single interface though, which indicates that proteins D and E bind to it competitively. Protein C interacts with proteins D and E differently, possibly binding multiple protein E molecules concurrently. In the interface-interaction network, all five proteins have recognizably distinct interaction behaviour. For clarity, the interface-interaction network is shown as a hypergraph rather than a graph.

However, isolating PPIs to these submolecular interfaces, which can be simply short subsequences but also complex structures (see Figure 1-2) formed from different sections of the polypeptide, is very difficult. Hence while networks that capture the specific interfaces in PPIs, called interface-interaction networks (IINs),[44] would more precisely model the physical interactions between proteins, they are no substitute for PPINs due to the sparsity of data. Some experimentally mapped interface-interaction data sets have recently become available, such as a set of PPIs mediated by SH3 domains in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*.[46,49,50] SH3 domains are peptide-recognition modules that bind to short linear peptides with characteristic proline-rich motifs. In graph form, the resulting SH3-mediated IINs are bipartite, though this may not be generally the case with other interface types (see Figure 1-3). Due to this bipartite property, certain network topology motifs, such as cliques, are absent while others, such as 4-cycles, are highly enriched, and analytical techniques designed for use on PPINs may not work with IINs.

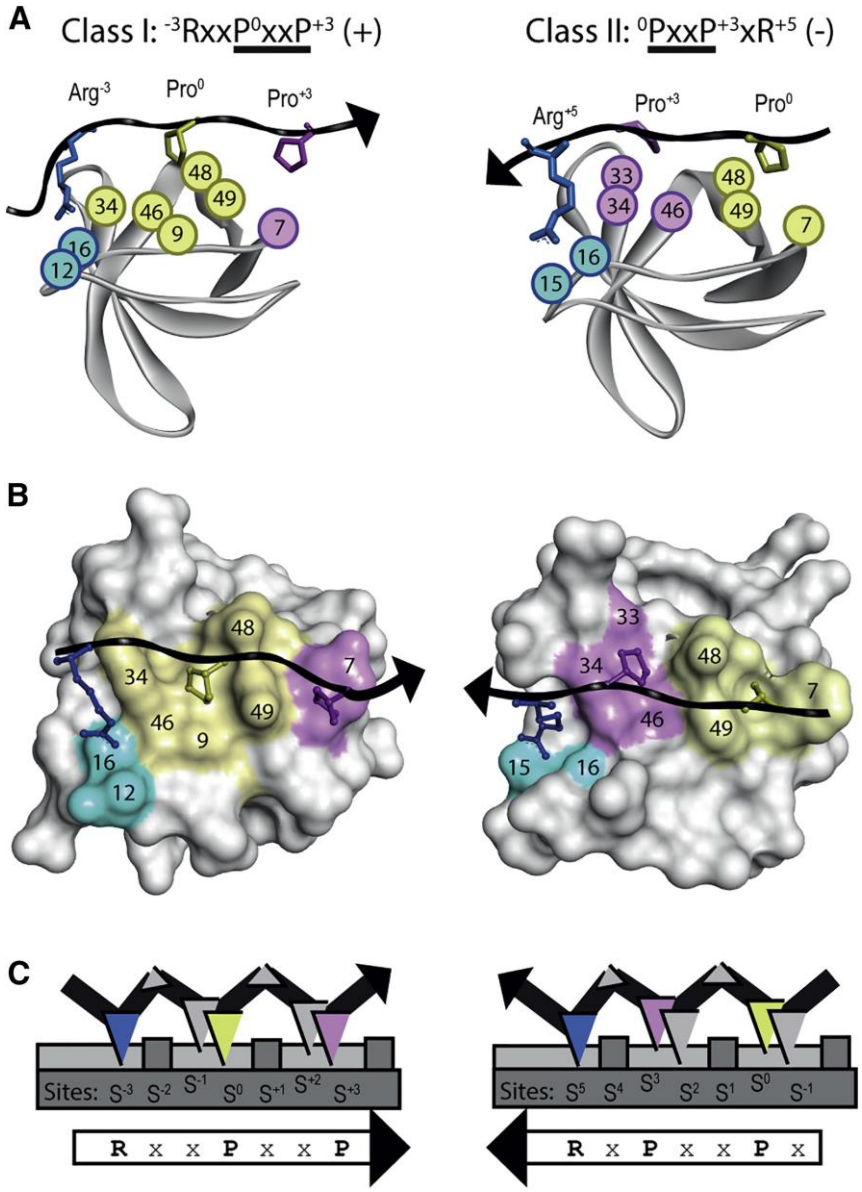


Figure 1-2 – Figure 1 from *Comprehensive Analysis of the Human SH3 Domain Family Reveals a Wide Variety of Non-canonical Specificities*.[50] Depicted are the canonical structures for protein-protein interactions mediated by SH3 domains. The SH3 domain is a complex, 3-dimensional structure consisting of approximately 60 amino acids, while its target ligand, shown as a black arrow in (A) and (B), is a one-dimensional peptide chain of approximately 8 amino acids. (Reproduced with permission.) **Original caption:** *(A) Representative crystal structures of SH3 domains in complex with class I (left, CTTN-1/1, PDB: 2D1X) or class II (right, CD2AP-1/1, PDB: 3U23) peptides. The peptide backbone is shown as a black tube with the C-terminus indicated by an arrowhead and side chains shown as colored sticks, as follows: Pro^{0} (yellow), Pro^{+3} (pink), Arg^{-3} in class I or Arg^{+5} in class II (blue). The SH3 domain backbone is shown as a gray ribbon and the residues that interact with the peptide are represented as spheres numbered according to the SH3 domain family alignment (Table S1) and colored to match the peptide residue that they contact. (B) Surface representation of the SH3 domains colored as in (A). (C) Schematic depiction of class I and class II peptide recognition. Peptide residues are depicted as triangles colored as in (A). The SH3 domain sites are depicted as gray boxes numbered accordingly.*

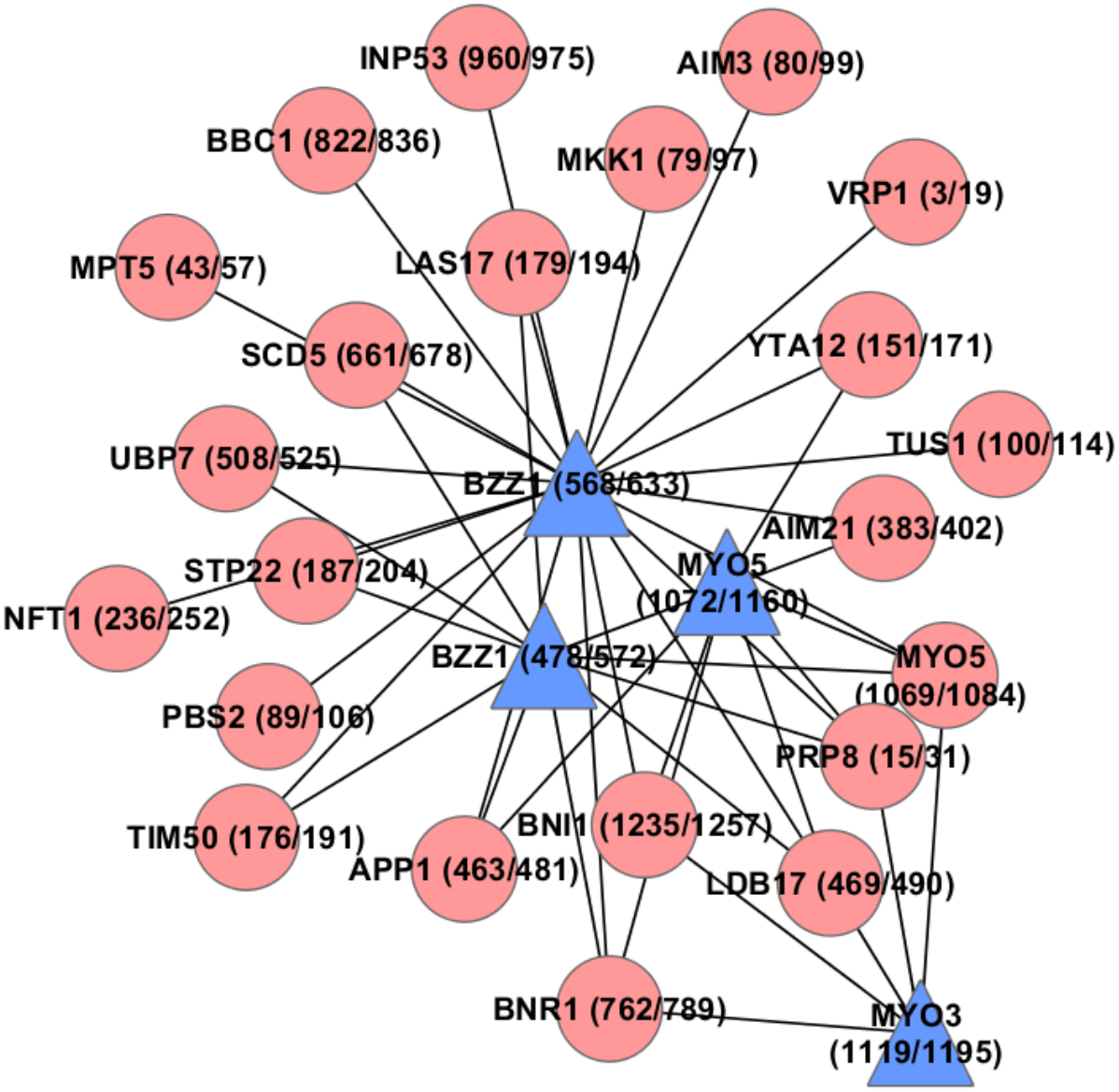


Figure 1-3 – A representative subnetwork of the S. cerevisiae SH3-mediated interface interaction network. Depicted in blue are four SH3 domains from BZZ1, MYO3, and MYO5; in pink are their binding sites, with the start and stop positions of their protein subsequences in parentheses. While SH3 domains can target the same site, they are not known to bind to each other, creating a bipartite network structure. Furthermore, since most SH3 domains bind to many binding sites, this creates a hub-and-spoke network topology wherein high-degree SH3 domain nodes occupy central hub positions surrounded by lower-degree binding site nodes.

1.2 Protein-Protein Interaction Network Alignment

One emerging new application for protein-protein interaction data is *network alignment*. Specifically, network alignment typically refers to the alignment of the PPINs generated from the whole or partial interactomes of different species, though the alignment of other networks is also

possible, such as metabolic networks or gene coexpression networks.[51-54] Network alignment is explicitly a tool for assessing comparative systems evolution via direct comparison of PPINs to identify conserved and divergent elements.

In its most basic form, given two PPINs, $G_1 = \{V_1, E_1\}$ and $G_2 = \{V_2, E_2\}$, wherein each node represents a protein and each edge a PPI, an alignment between G_1 and G_2 is some injective mapping between $\{(u_1 \in V_1, v_1 \in V_2), (u_2 \in V_1, v_2 \in V_2), ..., (u_n \in V_1, v_n \in V_2)\}$, where $n \leq |V_1|$ and $n \leq |V_2|$.[55] This formulation greatly simplifies the network alignment problem, though there are still O($n!$) possible alignments to consider. There are many variations on this formulation in the published literature. Among them are alignments of multiple networks, those with many-to-many equivalence classes rather than one-to-one injective mappings, and those that allow proteins to be aligned to multiple partners. These distinct formulations are inspired by various data, algorithmic, and biological considerations, and there is no consensus on a superior or single correct formulation. As such, in reviewing the existing network alignment literature, an "alignment" will refer to this simplest formulation unless otherwise noted.

1.2.1 Theoretical Considerations

The premise of network alignment is that given that PPIs are functional in nature, their evolution must be functionally constrained, and so the network topology of PPINs must also be functionally constrained. Then, as with sequence alignment performed on genes/proteins, network alignment performed on the interactome should be able to reveal the evolutionary history of the interactome. For example, a pathway or complex that is highly conserved between species would indicate a very constrained, possibly essential, function. Alternately, if a region of a PPIN is highly *rewired* in another PPIN – that is, some proteins in one species have vastly different interaction partners from their orthologs in another species – this indicates a lack of functional conservation for those proteins, and perhaps even the evolution of new molecular functions or modules.

By general consensus, network alignment has broadly been divided into two sub-problems, based on the expected output. This division is inspired by the distinction between local and global sequence alignment. In *global network alignment* (GNA), the inputs are two (or more) large networks, prototypically the whole interactomes of different species, and the output is an alignment between most, if not all, of the nodes of those networks.[55] In contrast, in *local network*

alignment (LNA), there is no expectation that the output alignment will cover most of an interactome. Instead, LNA accepts as input one or more large PPINs, plus possibly a small query network, and produces multiple small alignments of subnetworks that may overlap.

There are several established problems in computer science theory that resemble the network alignment problem. In the trivial case of aligning a network to itself, network alignment is akin to the graph isomorphism problem, which is of unresolved complexity, though a quasi-polynomial algorithm has recently been proposed.[56,57] In the case of aligning a network with a subnetwork, as might occur when querying an interactome with a pathway or a complex, network alignment is akin to the subgraph isomorphism problem, which is NP-complete.[58] In practice, however, neither of these problems are realistic representations of the biological considerations of network alignment; generally, there is no expectation that a meaningful exact (sub-)network match will be found, due to network evolution and rewiring. It is precisely because exact matches are not expected that network alignment holds relevance; if PPINs were fixed in the absence of evolutionary dynamics, the network alignment problem would simply be the problem of "aligning" proteins, already largely solved by sequence alignment.

Given that network rewiring results in distinct networks that should nevertheless be alignable, network alignment may best be considered a specific case of the inexact graph matching problem.[59] However, the biological characteristics of PPINs and our interest in the evolution of networks establish network alignment as a unique problem in need of novel methods, though prior inexact graph matching methods may still prove useful.

1.2.2 Local Network Alignment

Within network alignment, local network alignment (LNA) refers to alignment methods that produce smaller subnetwork alignments, in contrast with global alignment which attempts to align whole interactome networks. With its smaller scale, local alignment is computationally simpler and conceptually grounded in biology, as the aligners aim to identify pathways, protein complexes, or other connected network modules of interest to biologists.

Created in 2003, the first popular network alignment algorithm, PathBLAST, is a local alignment method that intends to identify pathways in a PPIN by alignment.[60] Taking a query, which may either be a pathway of interest or a larger PPIN, and a target interactome, PathBLAST merges

the two networks into a global alignment graph, based on sequence similarity between the protein nodes as determined by BLAST (nodes merged if E-value < 10^{-2}). Paths through this global alignment graph are then scored, with a composite Bayesian scoring function that attempts to assess the likelihood of true protein homology and true PPI identification,[61] and high-scoring paths are returned as aligned pathways. The scoring function also includes consideration for gaps or mismatches in the aligned pathways, penalizing their score but still allowing for alignment if the overall score is sufficient. A later iteration, NetworkBLAST, also aligned protein clusters or complexes using a seed-and-extend algorithm, which builds seeds of three or four nodes for each node in the network, and then uses a local search heuristic, adding high-scoring neighbours and removing low-scoring members, to expand the seed to up to 15 proteins (not higher, due to computational limitations).[62]

PathBLAST and NetworkBLAST established many of the characteristic features of LNA methods, such as the use of a seed-and-extend algorithm, the use of a global alignment graph, and a statistical scoring function based on a network evolution model. Though how each of these components and how they are assembled vary from method to method, most LNA algorithms retain these key features, distinguishing them from GNA methods. NetAligner adds additional edges to the global alignment graph, filling in edge gaps/mismatches based on the assumption that interacting proteins evolve at similar rates.[63] MaWISh includes duplication as a third evolutionary event to the model underlying its scoring scheme.[64] Graemlin uses a scoring matrix to allow searches for arbitrary structures.[65] AlignNemo uses connected 4-subgraphs as seeds to initialize the seed-and-extend algorithm, which then connects seeds together to form an alignment.[66] AlignMCL uses a Markov clustering algorithm to form an alignment from the global alignment graph.[67] GASOLINE uses bootstrapping to stochastically, iteratively extend seeds, and is capable of aligning multiple networks simultaneously.[68]

In recent years, however, there has been a decline in interest in LNA research, with GASOLINE being the last prominent LNA method published in the last five years. The attention of the network alignment community has largely shifted to focus on global network alignment instead.

1.2.3 Global Network Alignment

In global network alignment (GNA), the objective is to "find the best overall alignment between the input networks," as described in the paper describing, IsoRank, the first popular GNA

method.[55] IsoRank formulated the GNA problem as a spectral problem, creating an eigenvalue equation and then using an iterative algorithm to solve for the principal eigenvector. The eigenvalue equation used by IsoRank, like that used by PathBLAST, consists of two components added together: a protein sequence similarity component, E-values as computed by BLAST, and a network topology similarity component, in which the similarity score of every pair of nodes is partially distributed to their neighbours, in a PageRank-like manner.[69] Once an eigenvector is converged upon, the top scoring node pairs within that eigenvector are extracted and greedily aligned.

As PathBLAST did for LNA and network alignment in general, IsoRank established a key trend that persists in GNA research to this day, a function used to compute node similarity that is then used to guide alignment. In particular, many GNA methods continue use a similarity function that adds the two components, with a parameter used to control their relative weights, and BLAST to produce the scores for the protein similarity component. Unlike with PathBLAST and other LNA methods, the weighting of this scoring function is not derived from statistical analysis, but instead manually set to an arbitrary value, or for unspecified reasons.

In contrast to LNA, GNA methods are more varied, with many variations on the network topology measures used in the node scoring function and the algorithm used to transform that scoring function into an alignment. There are four major types of algorithms used in GNA methods: seed-and-extend, optimization, spectral, and genetic.

GNA seed-and-extend algorithms are highly similar to those used in LNA methods, beginning with a small, high-confidence seed alignment and then iteratively growing the alignment outwards. Many GNA seed-and-extend algorithms only extend along aligned edges, unlike LNA algorithms which account for interaction gaps and mismatches, which improves their performance on various network topology alignment metrics (see below). One common variation is to align in "shells" outwards from the seed, iteratively aligning all nodes at the same distance from the seed, then using some procedure to align unaligned "orphan" nodes. One such GNA method is GRAAL, which introduced the graphlet degree signature as a measure of the network topology similarity of nodes for use in node similarity scoring function, used throughout the GRAAL family of network alignment methods.[70] Two of GRAAL's descendants, C-GRAAL and MI-GRAAL, also use similar seed-and-extend approaches.[71,72]

Another common variation seen in GNA seed-and-extend algorithms is to use multiple seeds and mesh together the separately seeded alignments. By starting with one or more high-scoring seeds, seed-and-extend algorithms ensure that regions of high similarity are aligned, not sacrificed for overall alignment quality, and that the algorithm does not entrap itself with a poor initial seed choice. GHOST is one such method, similar to GRAAL.[73] GHOST uses a spectral signature in its node similarity scoring function, treats each shell to be aligned as a quadratic assignment problem, and adds a local search phase at the end to improve the initial alignment. IGLOO uniquely uses local network alignments as its seeds, and then performs global alignments with the remainder of the network.[74] NETAL and HubAlign are other examples of GNA seed-and-extend algorithms.[75,76]

GNA optimization algorithms create alignments by finding the alignment with the overall maximum of some optimized value. In contrast with seed-and-extend algorithms, optimization algorithms pay less attention to contiguity of aligned regions and may sacrifice regions of high similarity if it improves overall alignment. H-GRAAL is an example, using the Hungarian algorithm to find the alignment with the maximum sum of node similarity scores between aligned nodes.[77] PISwap is a unique method that uses local search to refine existing alignments, iteratively testing minor swaps in the alignment to optimize an evaluation function, although it has also been used to create de novo alignments.[78] The alignment algorithm in ModuleAlign uses the Hungarian algorithm to optimally align proteins based on a novel clustering-based similarity function, then uses local search to adjust the alignment to maximize alignment of edges.[79] SANA uses the simulated annealing heuristic to quickly converge to the optimal alignment.[80]

There are also several network alignment methods that employ genetic algorithms to seek optimally scoring alignments, such as Optnetalign, which uses a multiobjective memetic algorithm to optimize on both network topology and biological objectives concurrently.[81] Other network alignment methods using genetic algorithms include GEDEVO and MAGNA++,[82,83] as well as their multiple-network alignment versions GEDEVO-M and multiMAGNA++.[84,85]

GNA spectral algorithms cast network alignment problem as a linear algebra problem, and employ spectral methods to find an alignment, often to achieve greater speed. IsoRank is such an example, as is IsoRankN, a descendent designed for aligning multiple networks, and L-GRAAL, from the GRAAL family of network alignment methods, which uses Lagrangian relaxation.[86,87]

GHOST uses Lagrangian relaxation to solve a quadratic assignment problem to align subsets of proteins previously created using a seed-and-extend method.[73] GA, Natalie, and Fuse are all network alignment methods that use spectral algorithms to generate alignments.[88-90]

1.2.4 Assessment and Evaluation of Network Alignments

1.2.4.1 Biological Assessment

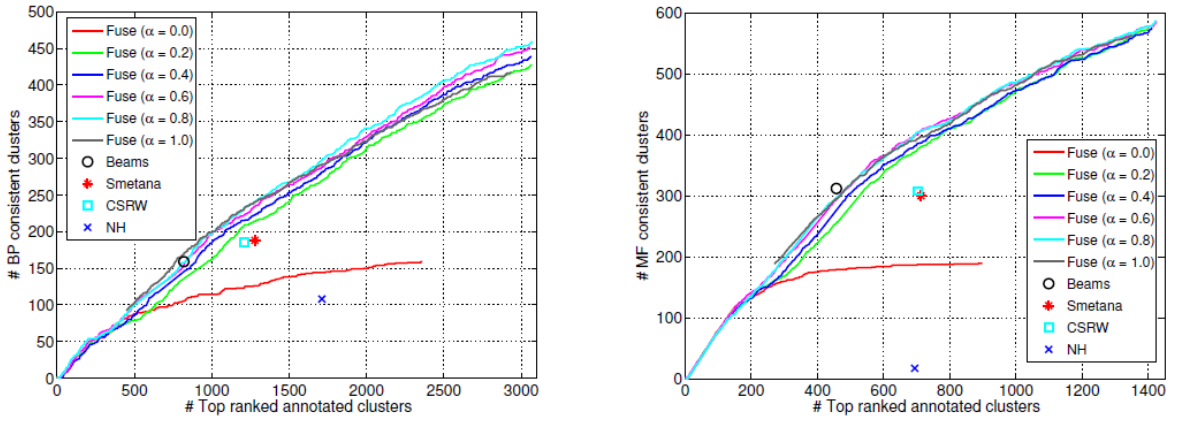
The evaluation of network alignment methods is commonly undertaken in two ways, biological and network-topological. The original application for GNA, as expressed by Singh et al., was to enable "species-level comparisons" and to detect functional orthologs, defined by Bandyopadhyay et al. as orthologous proteins that "play functionally equivalent roles."[55,91] To this end, Singh et al. analyzed the functional coherence of the yeast (*S. cervisiae*) and fruitfly (*D. melanogaster*) proteins aligned by IsoRank using functional annotations from the Gene Ontology (GO) Consortium, which maintains GO, a hierarchical ontology for protein function terms, and the associated GO database, a database of annotations attributing functional terms to genes.[92]

However, the method employed involved mapping all GO terms to ancestral terms at depth five in GO, discarding any terms at lower depths, and then counting the number of shared depth five GO terms. This methodology is advised against by the GO Consortium, as GO terms can occupy multiple levels in the hierarchy and there is no fixed notion of the significance of any particular level across GO.[93,94] For example, while both are second-level terms, "signaling" has more than 100 000 protein annotations in GO whereas "cell killing" has less than 2000,[95] meaning that alignment of two "signaling" proteins is not particularly significant compared to the alignment of two "cell killing" proteins. Furthermore, GO is irregularly shaped, with nodes and edges unevenly distributed across its branches, and GO annotations themselves are also unevenly distributed, making simple count-and-compare assessment methods unsuitable for determining the significance of two aligned proteins sharing an annotation.[96,97]

Unfortunately, due to the prominence of IsoRank, similar count-and-compare methods have regularly been used in assessing the quality of network alignments, including some that simply count the number of GO terms from all levels shared between aligned proteins, an inappropriately coarse measure of functional similarity.[75,76] It is only in recent years that the community has shifted towards more appropriate methods, such as semantic similarity or shared

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (KP) annotations,[98] though this shift has not been fully adopted.

Notwithstanding these technical quantitative issues, there is a qualitative conundrum for biological assessment of alignments based on existing functional annotation databases. Most network aligners include BLAST sequence similarity as an input, whether as E-values or bit scores. This creates a problem of circularity and/or redundancy, as many functional annotations and the experiments conducted to determine them are dependent on BLAST or other sequence alignment tools. As such, the high biological quality performance scores of some network aligners may be an undue result of sequence similarity input data dominating network topology considerations. For example, Fuse uses an α parameter to control the relative weights of their novel non-negative matrix trifactorization-based similarity score, which fuses together sequence similarity and network wiring patterns, and BLAST E-values.[90] However, the biological quality of the alignments generated using α values between $\alpha = 0.2$ (minimal BLAST E-value weight) and $\alpha = 1.0$ (total dependence on BLAST E-values) were nearly indistinguishable, whereas there was a steep drop-off in the absence of BLAST E-values ($\alpha = 0.0$) (see Figure 1-4). This suggests that Fuse's high-quality alignment results are highly dependent on BLAST E-values, and that perhaps the network-based considerations involved were not major contributors to the biological quality of Fuse's alignments.



Supplementary Figure 3. Functional consistency of Fuse's clusters for different values of α. For Fuse's protein clusters with different values of α, we plot the cumulative number of annotated clusters containing proteins from all 5 species (x-axis) against the number of them that are functionally consistent (y-axis). Clusters are ranked according to the scores computed as the sum of association scores between their proteins. We also report the total number of annotated and consistent clusters for the competing aligners (points in the plots). Biological process (BP) and molecular function (MF) annotations are considered separately in the left-hand-side and right-hand-side panels, respectively.

Figure 1-4 – Supplementary Figure 3 from *Fuse: multiple network alignment via data fusion*, including original caption.[90] Note that there is no meaningful difference between results achieved with α values between 0.2 or 1.0. With α = 1.0, Fuse uses only BLAST similarity without any network topology in its similarity score, suggesting that network information plays only a marginal role in Fuse's results. (Reproduced with permission.)

1.2.4.2 Network Topology Assessment

Network alignment assessment via network topology measures was not undertaken in the original IsoRank paper nor in LNA papers, but was popularized for GNA by the Pržulj lab's work on GRAAL. GRAAL utilized the measures edge coverage* (EC), which counts the number of edges aligned, and the largest contiguous connected subcomponent (LCCS), which attempts to measure the contiguity of the network alignment.[70] Additional measures such as induced conserved sub-structure (ICS) and symmetric substructure score (S^3), have since been introduced as refined EC measures.[73,99]

* Edge coverage is often called "edge correctness" in network alignment research, but this is a misnomer as it is simply a count of the edges aligned, for which there is no demonstrable correctness to their alignment.[80]

The connection between network topology measures such as EC, LCCS, ICS, and S^3 and the biological significance of network alignments has never been effectively established, however. Recent work has shown a negative or no correlation between maximizing these network topology measures and improvements on biological quality measures. When MAGNA optimized its alignments on EC, ICS, and S^3, on *S. cerevisiae-H.sapiens* and *C. jejeuni-E. coli* PPIN pairings, GO term enrichment showed no consistent improvement, either on the number of GO terms shared by aligned proteins or GO semantic similarity of aligned proteins[99] (see Figure 1-5 and Figure 1-6). Malod-Dognin et al. have shown that there is a trade-off between the S^3 and KP scores of alignments produced by L-GRAAL, HubAlign, Natalie, and MAGNA[100] (see Figure 1-7). Using OptNetAlign, Clark and Kalita also found either negative and no correlation between biological measures, the number of shared GO terms and BLAST bit-score, and the network topology measures, EC, ICS, and S^3 [81] (see Figure 1-8).

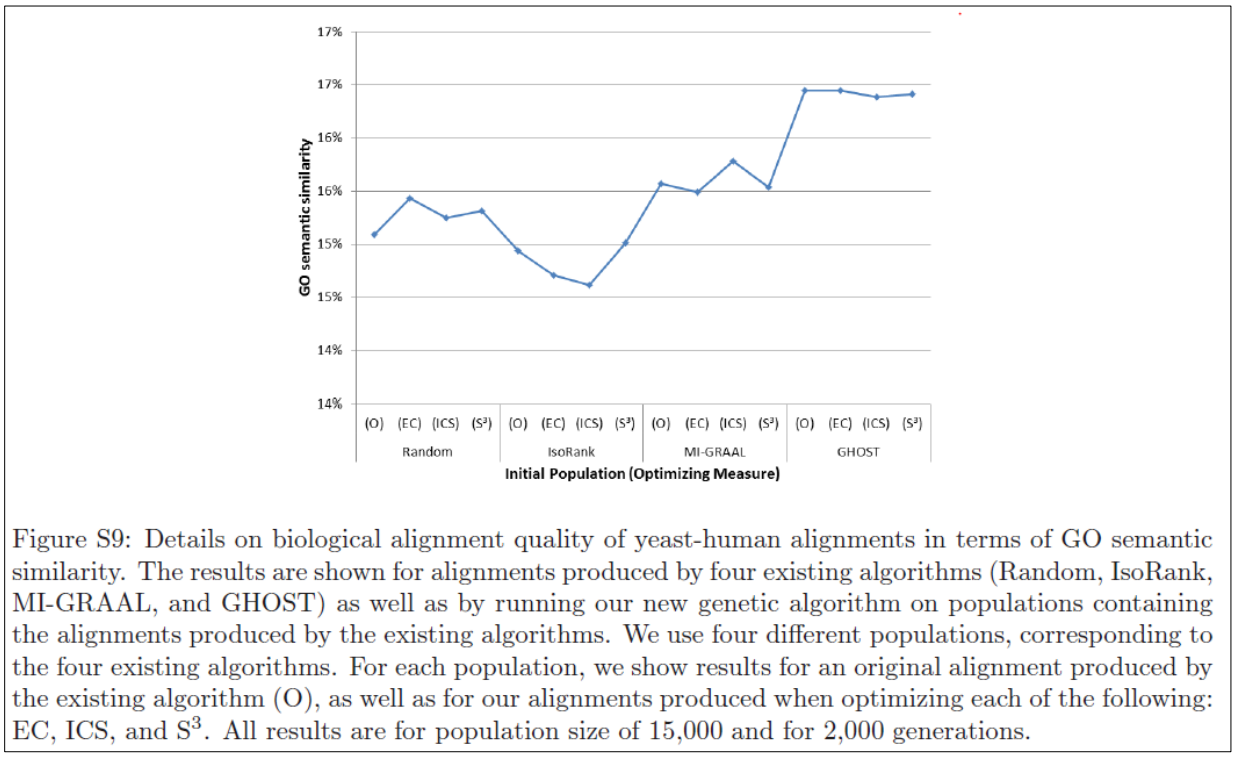


Figure S9: Details on biological alignment quality of yeast-human alignments in terms of GO semantic similarity. The results are shown for alignments produced by four existing algorithms (Random, IsoRank, MI-GRAAL, and GHOST) as well as by running our new genetic algorithm on populations containing the alignments produced by the existing algorithms. We use four different populations, corresponding to the four existing algorithms. For each population, we show results for an original alignment produced by the existing algorithm (O), as well as for our alignments produced when optimizing each of the following: EC, ICS, and S^3. All results are for population size of 15,000 and for 2,000 generations.

Figure 1-5 – Supplementary Figure S9 from *MAGNA: Maximizing Accuracy in Global Network Alignment*, including original caption.[99] MAGNA was used to optimize PPIN alignments of *S. cerevisiae* and *H. sapiens* generated by different alignment algorithms, using one of three network topology similarity measures (EC, ICS, and S^3). There was no pattern of increased GO semantic similarity of the aligned proteins found in the post-optimization alignments compared to the original alignments (O), suggesting that these three network topology similarity measures are a poor fit for predicting function and orthology, and for measuring alignment quality. (Reproduced with permission.)

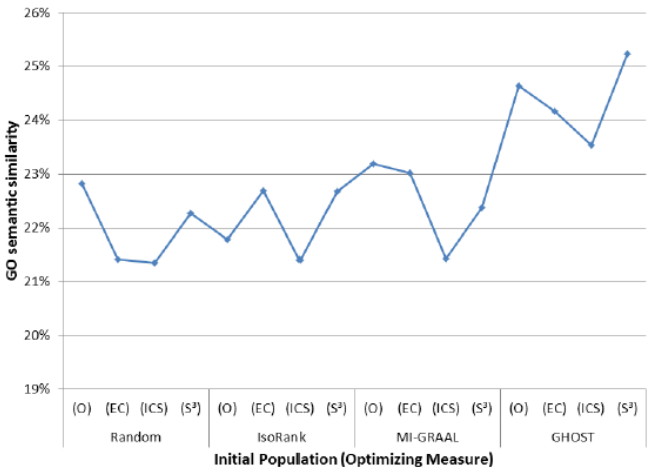


Figure S13: Details on biological alignment quality of *C. jejuni-E. coli* alignments in terms of GO semantic similarity. The results are shown for alignments produced by four existing algorithms (Random, IsoRank, MI-GRAAL, and GHOST) as well as by running our new genetic algorithm on populations containing the alignments produced by the existing algorithms. We use four different populations, corresponding to the four existing algorithms. For each population, we show results for an original alignment produced by the existing algorithm (O), as well as for our alignments produced when optimizing each of the following: EC, ICS, and S^3. All results are for population size of 15,000 and for 2,000 generations.

Figure 1-6 – Supplementary Figure S13 from *MAGNA: Maximizing Accuracy in Global Network Alignment,* including original caption.[99] MAGNA was used to optimize PPIN alignments of *C. jejuni* and *E. coli* generated by different alignment algorithms, IsoRank,[55] MI-GRAAL,[87] and GHOST,[73] using one of three network topology similarity measures (EC, ICS, and S^3). There was no pattern of increased GO semantic similarity of the aligned proteins found in the post-optimization alignments compared to the original alignments (O), suggesting that these three network topology similarity measures are poor fits for predicting function and orthology, and for measuring alignment quality. (Reproduced with permission.)

These results suggest a disconnect between the network topology measures used to evaluate network alignments, and the proposed objectives of network alignment to predict protein function and identify functionally conserved proteins. Maximizing network aligner performance on both network topology and biological quality measures are contrary objectives, leaving open the question of *how* aligners should align.

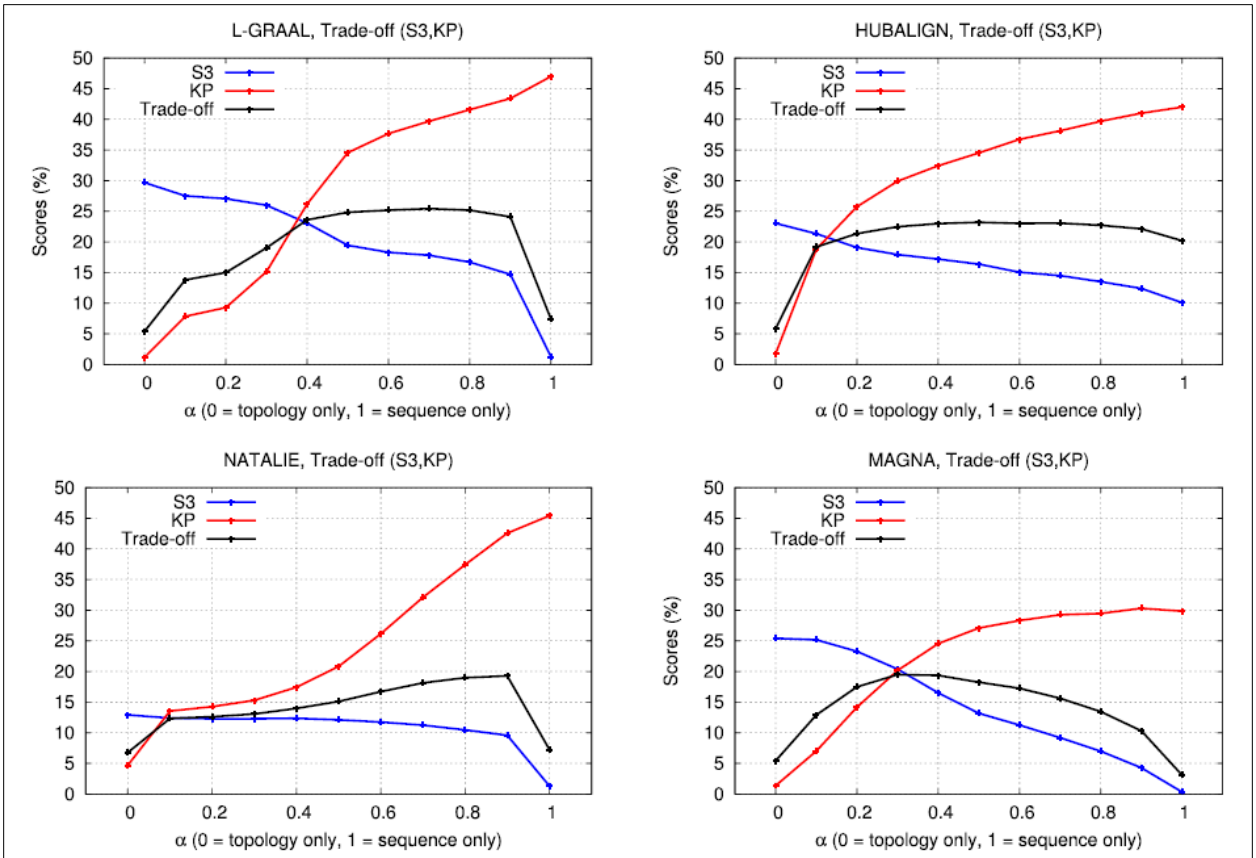


Figure 4. Effects of using topology and sequence homology. For L-GRAAL (top-left panel), HUBALIGN (top-right panel), NATALIE (bottom-left panel) and MAGNA (bottom-right panel), we report the average quality of the alignments that they produce on our 16 PPI network pairs, when aligning PPI networks by using topological information only ($\alpha = 0$, on the left on the x-axis), sequence similarity only ($\alpha = 1$, on the right on the x-axis) and their combinations (all α in between in increments of 0.1). The qualities of the alignments are measured with topological score S^3 (blue curves), biological score KP (red curves) and the trade-off score between biological and topological quality (black curves).

Figure 1-7 – Figure 4 from *Unified Alignment of Protein-Protein Interaction Networks*, including original caption.[100] Malod-Dognin et al. used four different network aligners (L-GRAAL,[87] HubAlign,[76] Natalie,[89] and MAGNA[83]) and adjusted their α parameters to control the relative weight of sequence (BLAST) or network topology similarity inputs. Their results show an explicit trade-off in the network topology evaluation measure S^3 and the biological quality evaluation measure KP (the % of aligned proteins sharing KEGG pathways) for the resulting network alignments. (Reproduced under the Creative Commons Attribution 4.0 International License.)

Table 1.

Correlation matrix for various objectives on the *S.cerevisiae* to *H.sapiens* alignment problem

	ICS	S^3	EC	GOC	Bit score sum	LCCS size
ICS	1.0	−0.132	−0.338	0.079	−0.103	−0.42
S^3		1.0	0.957	−0.964	−0.848	0.939
EC			1.0	−0.948	−0.796	0.987
GOC				1.0	0.81	−0.91
Bit score sum					1.0	−0.759
LCCS size						1.0

Unlike the other objectives, the number of nodes in the largest common connected subgraph (LCCS) was not set as an optimization objective in our program; it was computed afterwards

Figure 1-8 – Table 1 from *A comparison of algorithms for the pairwise alignment of biological networks*, including original caption.[81] Clark et al. used OptNetAlign to generate many non-dominated alignments of the S. cerevisiae and H. sapiens PPINs, optimizing on three network topology measures (ICS, S^3, and EC) and two biological quality measures (Gene Ontology Consistency and BLAST bit-score). They examined these 571 alignments and found the above correlations of the five measures, plus LCCS, in those alignments. The two biological quality measures (GOC and BLAST bit-score) showed no or strongly negative correlation with the three network topology measures, again indicating that network topology similarity is not an appropriate measure of an alignment's biological quality nor its ability to make biological conclusions. Furthermore, the high correlation between GOC and BLAST bit-score hints at the confirmation bias problem in using BLAST bit-score as an input to network alignment while also using GO term similarity as an evaluation metric. (Reproduced with permission.)

1.2.4.3 Simulation-based Assessment

Some network alignment methods are also evaluated based on their performance in simulated trials. The advantage of aligning simulated networks is that the correct node pairings are known in their totality beforehand, without any confirmation bias introduced by the use of other sources of biological information, allowing for direct assessment of an alignment method's ability to correctly align nodes (proteins). Evaluation using fully simulated networks is rare, due to continued debate on how best to simulate PPINs (see 1.3.2 Models for Protein-Protein Interaction Networks).

Instead, researchers often take existing PPINs, randomly rewiring edges to simulate network evolution and PPI noise, then align the rewired networks with each other and/or the original PPIN.[70,75,99] Additionally, several methods use an *S. cerevisiae* PPI dataset from Collins et al.,[101] adding lower confidence PPIs to the *S. cerevisiae* PPIN before aligning the two.[73,77,99] The usefulness of these simulations, however, depends upon the verisimilitude of their approximation of real PPIN variation. Uniform random removal and addition of edges to PPINs is very unlikely to faithfully approximate evolutionary divergence in real-world interactomes. Furthermore, both simulation methods, as typically employed in the literature, use inappropriately low rates of noise, with a maximum of only 25%, in contrast to higher observed rates of PPIN rewiring.[102-105]

1.2.4.4 Summary

Assessment of network alignment methods and the alignments they generate remains a difficult problem to unpack, with various issues that trouble all popular metrics. Whether biology-, network topology-, or simulation-based, all the popular evaluation metrics present difficult conundrums for network alignment researchers. Biological assessment metrics have problems with circularity and possibly overvaluing naturally dominating biological input data, such as protein sequence alignment. Network topology metrics may not be measuring the alignments' biological informativeness nor their ability to produce biologically meaningful insights. Simulation-based metrics may be inaccurate due to our poor understanding of PPINs and PPIN evolution.

It should be noted that network alignment, if developed into a successful, validated biological research tool, could answer these very questions that trouble its development. Network alignment could, in an ideal future, serve as an alternative tool for determining protein function and inspiring proteomic research, illuminate how PPINs encode protein function and reveal the underpinnings of PPIN evolution, and offer us insights on the structure of PPINs. Ultimately, the ability of extant network alignment approaches to answer these questions, however, is dubious. When Malod-Dognin et al. evaluated the biological quality of 8 different network aligners (Natalie, L-GRAAL, PiSwap, HubAlign, SPINAL,[106] MAGNA, ModuleAlign, and OptNetAlign) based on the number of aligned proteins that share GO terms, divided into the categories of biological process, cellular component, or molecular function terms, all the tested methods performed rather poorly[100] (see Figure 1-9). Fewer than half of the aligned proteins

shared *any* common GO terms in any of the above categories in 16 separate experiments with different PPIN pairs, and the best average performance for all aligners was less than 25%. If the primary objective of network alignment, in particular GNA, is to enable determinations of protein function, these results are insufficient and unpromising.

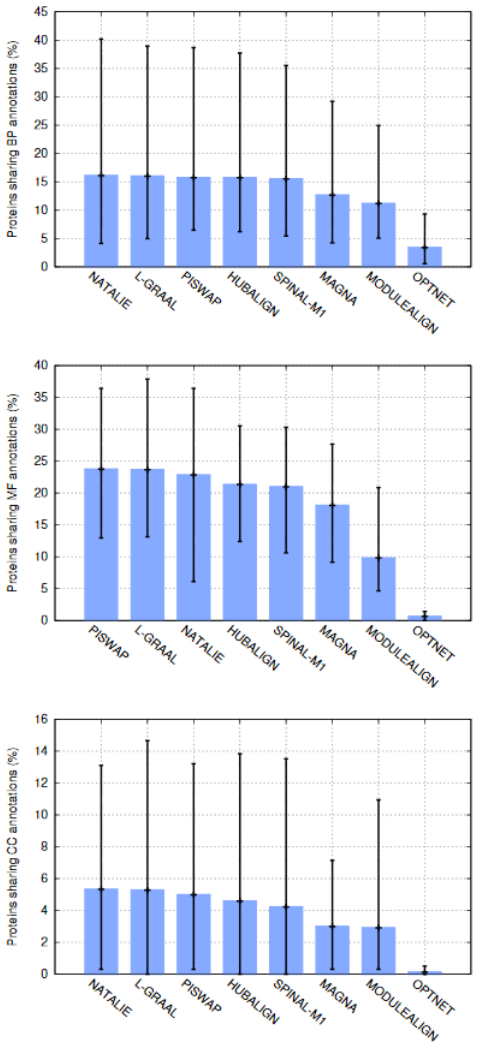


Figure 1-9 – Part of Supplementary Figure 2 from *Unified Alignment of Protein-Protein Interaction Networks*.[100] Three charts showing the performance of 8 different network aligners[76,78,79,81,87,89,99,106] in aligning proteins with at least 1 shared GO annotation, divided into biological process (BP), molecular function (MF), or cellular component (CC) GO terms. The coloured bar shows the average performance by the aligner on 16 different PPIN pairs, while the whiskers show minimum and maximum. Note that y-axes are not standardized, and that aligner performance is rather low, with the best average performance at ~16% for BP, ~24% for MF, and ~6% for CC, levels that are too low to consider for functional prediction. (Reproduced under the Creative Commons Attribution 4.0 International License.) **Original caption:** *Detailed performance comparisons. Network aligners (x-axis) are compared according to the topological and biological quality (y-axis) of their alignments. The error bars show the smallest, the average and the maximum of these scores over the 16 PPI network pairs, respectively. The left panels present the results for the topological scores (from top to bottom: NC, EC, ICS, S^3 and LCC) and the right panels present the results for the biological scores (from top to bottom: KP, GO-BP, GO-MF and GO-CC). In each panel, aligners are sorted from the best performing (on the left) to the worst performing method (on the right).*

1.3 Protein-Protein Interaction Network Evolution

Protein-protein interaction network alignment is often presented as a static problem, seeking the optimal alignment of two (or more) fixed input graphs, but underlying the problem is the dynamic process of network evolution, which creates the differences in the networks being aligned. As PPINs encode function, and evolution is both constrained and guided by function, the questions of how network alignment should be performed and how network evolution occurs are effectively intertwined. Better understanding of how function is encoded and conserved within the network would, for example, indicate how networks and proteins should be aligned. Similarly, better understanding of network divergence and positive selection in the network may indicate network regions that *should not be aligned*, an important yet largely unaddressed consideration that distinguishes local and global network alignment.[107]

While network alignment research remains in its infancy and primarily focused on simply identifying functionally similar protein pairings, it should eventually develop into a tool to investigate and understand network evolution, just as sequence alignment serves for genetic/genomic evolution. Considering network alignment against the larger background of network evolution will enable insights into the contours of the network alignment problem, which can then be applied to create network alignment and analysis tools that more precisely answer questions of scientific interest.

1.3.1 Fundamentals and Applications

As with the genome, the interactome is a potential treasure trove of information that could further our understanding of the cell. Within the PPINs that represent the interactome are all the protein complexes and pathways studied in molecular biology. These modules are critical components in our current understanding of cellular biology, but much of the interactome lies outside these well-studied modules, and thus also beyond our understanding. With the availability of interactomic data from many species, it is possible that we can leverage this data using comparative evolution to better understand the yet poorly studied parts of the interactome.

Even with the PPI data quality issues described earlier, many significant observations have been drawn from comparative network analysis, with various implications for our understanding of evolution. It has been found that genes under positive selection are more likely to be located at

the periphery of the PPIN.[108] Network topology has also been used to identify certain proteins that are particularly important to the cell, specifically hub and bottleneck proteins.

Hub proteins are high-degree proteins, and their network importance is reflected biologically by their tendency to be essential proteins.[45,109,110] Hub proteins can be divided into "date" hubs, proteins with many partners that compete for the same binding interface, and "party" hubs, proteins that bind to multiple proteins concurrently via multiple binding interfaces. Date hubs play an outsized role in the connectivity of the PPIN, while party hubs serve as a crux for functional proteomic units. Their distinct roles influences their sequence and structural evolution, with date hubs having higher levels of disorder to facilitate docking of different proteins in the same region.[111] Bottleneck proteins are proteins that occupy critical junctions between highly connected portions of the interactome, identified by their high betweenness centrality.[112] Like hub proteins, they tend to be essential, due to their role in linking different network modules, which are themselves arranged around hub proteins. These distinct meta-functional roles played by hub and bottleneck proteins may themselves exert distinct evolutionary forces upon these proteins.[113]

These examples demonstrate that there are complex dynamics between genetic evolution and network evolution; understanding the interplay between evolution on these different levels will be critical to developing a full understanding of the cell.[47] Unlike genes, however, of which there are many in every species, each species has only one interactome, as they are currently abstracted. This limits opportunities for inter-species comparison, elevating the importance of high-quality comparative network analysis in identifying the patterns within interactomes that would illuminate network evolutionary processes.

However, in addition to these modules, there is an extensive amount of noise, including experimental noise (i.e. false positives and false negatives) and also true PPIs that contain varying amounts of biological information that has not been understood yet. Distinguishing the true noise from the less understood parts of the interactome poses a major challenge, but comparative network analysis may be a potent tool in addressing it, by observing patterns of conservation across different interactomes.

One way this is done is by considering the interologs of different PPINs. An interolog is an analogue to orthologs for protein-protein interactions; if two interacting proteins in the same

species each have, in another species, orthologs that also interact, then those two interactions are interologs.[114] Interologs have been used to directly transfer PPIs, inferring PPIs in one PPIN from the presence of their interologs in the other, a process called interolog mapping.[18,112,115-117] This allows the transfer of information from well-studied interactomes to poorly studied interactomes, enabling analyses otherwise impossible due to a lack of data, such as analyzing the relationship between network rewiring and gene essentiality in *S. cerevisiae* and *M. musculus*.[21] Interologs, or the absence of them, can also be used to identify and study protein complexes and their evolution.[118,119]

There is, however, a significant body of work that suggests that our understanding of PPINs is insufficient, or the existing PPI data is insufficient, for such inter-species interactomic analyses. Some studies have indicated that interolog conservation between species is too low for interolog mapping as a general strategy.[104,105] There have also been studies questioning our statistical understanding of PPINs and, consequently, conclusions based on those studies, such as the dichotomy between date and party hubs and ability to predict function using PPINs.[120-122] Thus, as PPI data continues to be collected and used in new applications, work must continue on understanding the fundamentals of PPINs and PPIN evolution.

1.3.2 Models for Protein-Protein Interaction Networks

There is an ongoing debate in PPIN evolution research over the best network generation model to explain and/or simulate PPIN evolution. The development of a network generation model that could accurately simulate real-world PPINs would likely provide extensive insights into the mechanisms of network evolution. More practically though, accurate PPIN generating models could generate simulated PPINs to serve as a statistical background for network analyses or as a testbed for better evaluation of network alignment methods.

Current popular PPIN models fall into three general categories: scale-free (Barabási–Albert), geometric, and duplication-divergence models. Scale-free networks have a degree distribution that follows the power law, which supposedly makes scale-free models suitable for a variety of networks, including PPINs, the World Wide Web, and social networks.[123] The Barabási–Albert model generates networks using preferential attachment: nodes are added to the network one at a time, and edges drawn from each new node to existing nodes randomly, but with probabilities favouring high-degree nodes.

Geometric random graphs are, like scale-free networks, used to model non-biological networks, like social networks. To generate a geometric graph, nodes are randomly scattered across a geometric surface and edges drawn between nodes if the distance between them is less than a specified parameter. There has been evidence that geometric random graphs are better models for PPINs along with other statistical evidence suggesting that PPINs are not truly scale-free.[124] However, geometric random graph models do not simulate evolutionary processes in any manner, as all nodes and edges are added to the network simultaneously. In contrast, scale-free models remain popular for PPINs as preferential attachment represents a biologically feasible mechanism for network evolution, though preferential attachment cannot fully explain PPIN evolution alone.

Finally, the duplication-divergence (DD) model generates networks using node duplication as its principal mechanism.[125] The evolutionary mechanisms of duplication and divergence have been featured in evolutionary models for gene and domain evolution, in addition to PPINs.[126,127] Beginning with two connected nodes, nodes are randomly selected to be duplicated along with their interactions, and then interactions are randomly removed from one or both of the duplicates. Unlike the scale-free and geometric random models, this model is not commonly used for networks other than PPINs, as it fundamentally incorporates a mechanism, gene duplication and divergence, that is unique to biology. However, duplication-divergence remains a highly simplified model of network evolution; while it may simulate real PPINs statistically,[128] it does not specifically model other network events, such as de novo protein gain or interaction rewiring.[107,129,130] The basic DD model, however, can serve as a platform for more variations that incorporate additional evolutionary events.[131]

Generally, these models are evaluated based on their ability to simulate random *de novo* networks that bear similar graph theoretic statistics to real PPINs, and not on their ability to explain or generate testable hypotheses for individual proteins or clusters. Given that much molecular biology research is structured around individual proteins, complexes, and/or pathways, this has resulted in an information gap, as the global perspectives often adopted in PPIN research are incompatible with the local perspectives more common in molecular biology as a whole. The absence of PPIN models with explanatory power has, for example, driven researchers interested in PPIN alignment to formulate novel PPIN models to serve as foundation for their alignment methods,[64,132,133] but these models are crudely parameterized and underdeveloped. Superior,

more reliable PPIN models would enable development of network alignment tools with the precision and specificity needed to extract relevant biological insights from PPINs.

1.4 Protein-Protein Interaction Network Alignment Visualization

As increasing amounts of molecular biology data have been generated in the past three decades, there has been a corresponding increase in the number of visualization tools developed to facilitate human observation and assessment of this data. While in the beginning, generic all-purpose graph visualization tools were sufficient,[134,135] the flood of data available to molecular biologists, in particular data of different types, demanded the development of new tools to not just visualize, but also integrate and partially automate analysis of this data. The "hairball problem" in particular, wherein large networks appear as a dense, visually indecipherable cluster when visualized, necessitated tools that could quickly and efficiently prune irrelevant information and present to users tightly arranged views, which encouraged specialization amongst network visualization tools. These tools include Cytoscape,[6] NAViGaTOR,[136] and OrthoNets.[137]

However, network alignment visualization is quite limited. OrthoNets, a Cytoscape app, creates side-by-side visualizations of orthologous subnetworks, but it is rather rigid in its approach, focused specifically on orthologous proteins and their neighbours, and automatically retrieves data from fixed, now outdated, sources. With the many network alignment methods now available, and their shaky ability to reveal biologically meaningful relationships, biologists interested in using network alignments as an information source need the ability to quickly visualize and assess the relevant results. This is especially true for global network alignments, as they may produce suboptimal alignment results in the region of interest in a compromise for overall alignment quality. Global network alignments also contain an overwhelming amount of data from the alignment of two whole interactomes, effectively creating a "double hairball" problem.

Another Cytoscape app for network alignment visualization exists for the network alignment method GASOLINE[68] (see Figure 1-10). Somewhat confusingly, this app is also named GASOLINE.[138] GASOLINE the app is designed to execute GASOLINE the method on user-provided input files, and display the results. Like with OrthoNets, the rigid workflow integration in GASOLINE the app limits its utility for those interested in other alignment methods, and there

is little data integration capability to facilitate incorporation of GASOLINE into more extensive workflows.

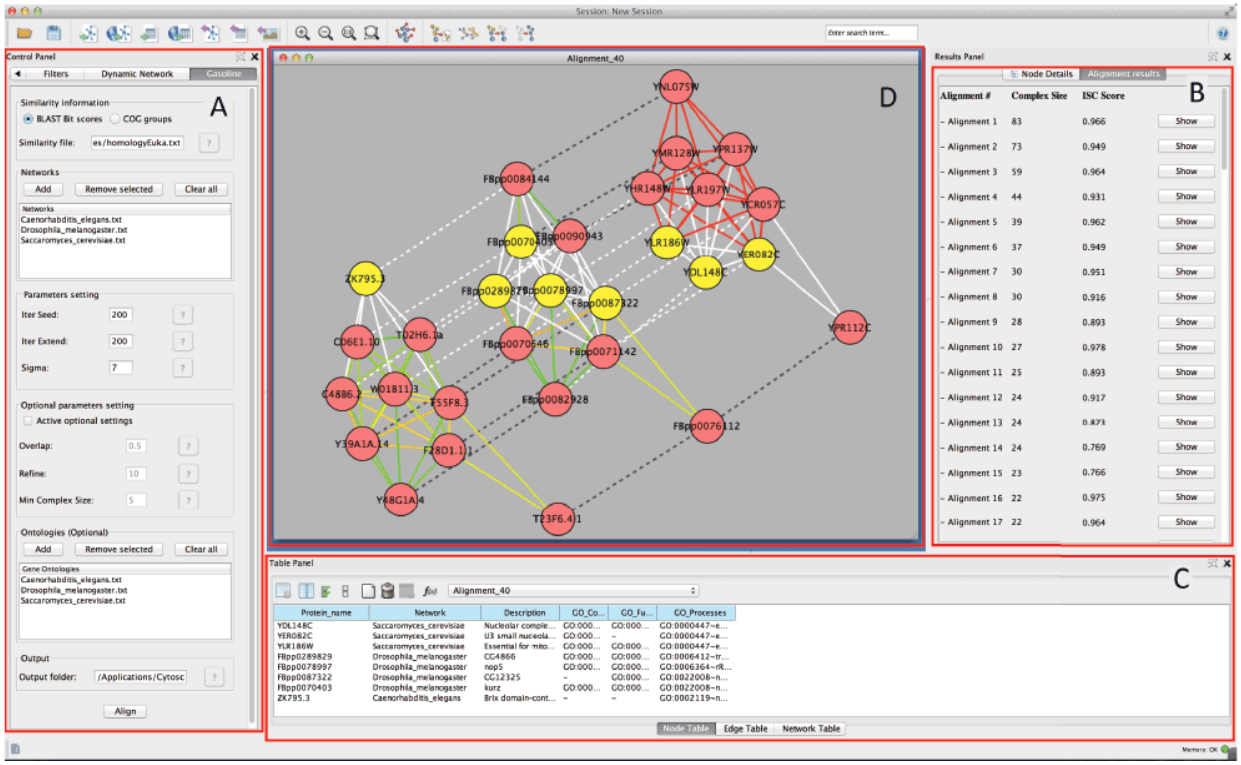


Figure 1-10 – Figure 1 from *GASOLINE: a Cytoscape app for multiple local alignment of PPI networks*.[138] The GASOLINE app is a tool to help users run the GASOLINE algorithm on their own input data and visualize the resulting local network alignments. (Reproduced under the Creative Commons Attribution 4.0 International License.) **Original caption:** ***A****) GASOLINE parameters;* ***B****) Alignment results;* ***C****) Description of selected proteins and associated GO terms;* ***D****) Alignment visualization: intra-edges are represented with solid lines and coloured according to their weight (green for low values, yellow for medium values and red for high values); inter-edges are drawn with dashed lines.*

Chapter 2

Alignment of Interface-Interaction Networks

This work was published in *Scientific Reports*, **5**:12074: Law, B., Bader, GD. (2015), GreedyPlus: An Algorithm for the Alignment of Interface Interaction Networks.

I conceived of, designed, and conducted this project. Gary D. Bader supervised and advised this project.

2 Alignment of Interface-Interaction Networks

2.1 Introduction

A major objective of biology is to understand how complex biological systems are assembled from their components into functional units and how they evolve. In molecular biology, efforts have increasingly focused on how proteins and other molecules interact, and determining how their interplay affects biological phenotypes, including disease. This has driven work in interactomics, as better, cheaper high-throughput methodologies allow us to systematically map the dynamic molecular interactions in a cell.[139] To aid the evolutionary study of these networks, a number of network alignment methods have been developed.[140]

Recently, protein interactions have been mapped at the level of amino acid positions, which can be represented as an interface-interaction network (IIN), where nodes represent binding sites, such as protein domains and short sequence motifs.[44,46,49,109,141] These networks provide a more accurate picture of how protein interaction networks are organized in biological systems. Thus, studying the function and evolution of these higher resolution networks should provide new biological insights. However, current protein interaction network alignment algorithms are not designed to align these networks and generally fail to do so in practice. In response, we developed GreedyPlus, the first algorithm designed to align IINs. In the next sections, we provide background about the network alignment problem, introduce IINs and review existing algorithms. We then describe the GreedyPlus algorithm and associated input data, comparisons with existing protein-protein interaction network alignment methods and results aligning IINs from different species.

2.1.1 Network Alignment Theory

In the trivial case of finding the ideal alignment of a network to itself, the network alignment problem is equivalent to the classic graph isomorphism problem, which is of unknown complexity[142]. However, as biological networks evolve, we expect divergence between the networks via the addition and deletion of both nodes and edges, and thus the objective of network alignment is to find similarity between networks rather than perfect isomorphisms. In the particular case where one network is a subnetwork of the other, the network alignment problem is specifically the subgraph isomorphism problem.[58] In the general case, the network

alignment problem degenerates into many instances of the subgraph isomorphism problem with loosened constraints; particularly, the objective is to find a set of non-overlapping, partial isomorphisms of all possible subnetworks of both networks. Given that the complete protein-protein interaction networks (PPINs) of species such as human and yeast[101,143] number in the thousands of nodes and edges, and that the subgraph isomorphism problem is NP-complete, an optimal computational approach is unfeasible and heuristics and approximations must be used.

2.1.2 Interface-Interaction Networks

Interface-interaction networks are a refinement of PPINs wherein proteins are subdivided into their separate interaction interfaces.[44] We choose to represent the IIN using a traditional graph model where a node represents a specific binding site and an edge represents a physical interaction between two binding sites on their respective proteins. The IIN is thus a higher resolution version of the PPIN.

The higher resolution of IINs allows for new biological insights that cannot be derived from standard PPINs. For example, IINs can distinguish between *date hubs* – proteins that interact with many partners, but at different times or in different locations - and *party hubs* – proteins that interact with many partners simultaneously.[45] While these distinct types of hub proteins will appear identically in a PPIN, in an IIN, the former will have few binding sites that are reused for many different interaction partners whereas the latter will have many binding sites that are specific for each interaction partner. This is useful to help elucidate the evolutionary processes and constraints acting on hub proteins. The study of IINs will also help interpret how domain and binding site gain and loss affect the PPIN, predict PPIN perturbations caused by sequence mutations that affect binding sites, and allow in-depth analysis of how protein-protein interactions are formed and lost.[46-48]

Network topology differences between IINs and PPINs, however, mean that algorithms designed to operate on PPINs may not function properly with IINs. While PPINs are often sparse, IINs are much more so, with each PPIN node (protein) split into multiple nodes that represent the different binding sites on that protein. Similarly, while PPINs exhibit a hub-and-spoke network topology, with many low-degree and fewer high-degree nodes, this characteristic is exaggerated in IINs. For example, protein-recognition modules, such as protein kinases or SH3 domains, are often capable of binding many different proteins, leading to relatively few high-degree nodes

connected to many low degree nodes. Additionally, due to binding specificity similarities, different domains will often recognize the same ligands, forming a multi-fan network topology. Methods that depend on the neighbourhoods of nodes having distinct local network topologies to generate their alignments get confused by these repeated patterns and thus perform inconsistently.

Experimentally mapped interface-interaction data across species have recently become available, such as a set of interactions mediated by SH3 domains in *Saccharomyces cerevisiae* (budding yeast)[49] and *Caenorhabditis elegans* (worm)[46] SH3 domains are peptide-recognition modules that bind to short linear peptides with characteristic proline-rich motifs. The resulting IINs are bipartite, though this may not be generally the case. Due to their bipartite property, certain network topology motifs, such as cliques, are absent while others, such as 4-cycles, are highly enriched. Existing PPIN alignment algorithms have not been designed for bipartite networks and can fail to align these networks. The graphlet degree signature similarity measure used by GRAAL,[70,144] for example, loses most of its resolution on a bipartite graph due to the absence of odd cycles. Alternatively, the bipartite nature of the networks confounds IsoRank,[55] as its node similarity measure can get stuck oscillating between domain and binding site nodes rather than converging.

To address the IIN alignment problem, we developed a new algorithm called GreedyPlus, which considers bipartite IINs by design.

2.1.3 Protein-Protein Interaction Network Alignment

Even though we argue that IIN alignment represents a different problem to PPIN alignment, the problems are related in their approach and we review PPIN alignment work here. Previous network alignment research has focused on protein-protein interaction networks, although other network types have been studied.[145] Previous PPIN alignment methods have sought to identify pairs of orthologous proteins and/or functionally orthologous proteins. Mirroring biological sequence alignment techniques, PPIN alignment methods have broadly taken two approaches: local alignment and global alignment. Local alignment algorithms seek small subnetworks that are similar in network topology, emphasizing regions of high-confidence alignment between the two networks. Typically, these methods use protein sequence alignment as a primary indicator of protein orthology, and then incorporate network information to identify clusters of sequence-

similar proteins; these clusters in the network, then, are considered putatively orthologous functional units.

PathBLAST,[60] one of the first published PPIN alignment methods, and its successor NetworkBlast[62] are examples of local network alignment algorithms. Both methods begin by identifying all pairs of proteins between the two input networks with significant sequence similarity (using BLAST E-values),[3] formulating each pair as nodes within a global alignment graph, and filling in the edges between these paired protein nodes using interaction data. In the global alignment graph, edges can be aligned (edges exist in both input networks), gapped (an edge exists in only one input network), or mismatched (no edge exists in either network), implying an abstract model of network evolution. A scoring model is then used to score the aligned proteins, and the high-scoring pairings are combined into a small pathway or complex as the final result.

Generally, the local network alignment strategy is similar to that for local sequence alignment, beginning with a seed that can be aligned with high confidence, which is often based on BLAST scores. A scoring scheme is defined, often based on an explicit evolutionary model, and then the alignment is extended outwards from the seed along network edges, incorporating as many other protein pairs as possible and optimizing on the score. NetAligner,[63] for example, assumes that interacting proteins evolve at similar rates as part of scoring edge mismatches and gaps. MaWISH[64] formulates an evolutionary model consisting of three events: match, mismatch, and duplication, which are used to develop a scoring scheme for optimization and thresholding. The explicit use of an evolutionary model to generate a scoring scheme is not novel; as with sequence alignment, any network alignment method implies an evolutionary model. However, as protein-protein interaction network evolution remains a largely mysterious process, the evolutionary models underlying the scoring schemes are diverse.

Otherwise, the local network alignment problem is well defined. The objective is to identify small, well-defined interactomic units – such as protein complexes or pathways – that are analogs within the input networks. However, by focusing on local regions, they may miss global aspects of network evolution. Additionally, as certain network topology patterns appear frequently in PPINs, such as cliques and hubs, local network alignments can improperly align subnetworks corresponding to these patterns. This is typically prevented using minimum

sequence similarity thresholds, explicitly or implicitly, to block the alignment of proteins with dissimilar sequences. As a result, these methods may miss functionally and topologically similar protein pairs that have dissimilar sequences.

Global network alignment methods attempt to align all or most of the proteins in two or more PPINs. These methods typically build interactome-wide alignments either by seeding an initial alignment and then extending it or by seeking a global optimum according to some scoring mechanism using methods such as the Hungarian[77] or the PageRank[55] algorithms. Global alignments likely have much higher false positive rates than local alignments as they align many more protein pairs, even those for which evidence is weak. Still, global alignment methods have produced network alignments with significant levels of functional similarity between aligned proteins.

The IsoRank algorithms – IsoRank[55] and its successor IsoRankN[86] - adopt a global approach to the PPIN alignment problem, formulating a set of mathematical equations and solving them concurrently across the entirety of the two networks, in a manner similar to the PageRank algorithm. The intuition behind the IsoRank algorithm is that two nodes should be aligned if their respective neighbours should be aligned, considering similarities of neighbours and BLAST sequence similarity. To solve for all possible node pairings, the problem is reframed as an eigenvector, and approximated using the power method. Once convergence is achieved, the nodes are aligned greedily based on their similarity scores. Neither network topology similarity nor an evolutionary model for networks are explicitly incorporated in this approach.

GRAAL[70] and H-GRAAL[77] focus on the use of graphlet degree signatures[144] as a purely network-topology-based measure of node similarity. GRAAL and the related MI-GRAAL[72] use a seed-and-extend approach aligning in expanding radii from the seed nodes in both input networks, aligning the nodes at each radius greedily. H-GRAAL, like IsoRank, formulates the global network alignment problem as a minimum-weight bipartite matching problem and solves this problem using the Hungarian algorithm. C-GRAAL[71] uses BLAST sequence similarity in a seed-and-extend approach where nodes with high neighbourhood densities are selected as seeds, greedily aligning their neighbourhoods, and then using a common neighbourhood mechanism to align further.

Alternatively, network alignment algorithms can use evolutionary models to score possible alignments in terms of likelihood, as BLAST does with sequence alignments. Unlike the IsoRank and GRAAL algorithms, Graemlin[65] and Graemlin 2.0[132] explicitly formulate a model for network evolution, consisting of four distinct evolutionary events for Graemlin and six for Graemlin 2.0. These models are trained on pre-existing protein orthologies from KEGG,[146] and then used to score potential alignments between networks. However, even using a seed-and-extend method that takes an iterative approach to alignment creation, the number of possible events results in an exponential number of possible steps at each iteration, requiring complicated heuristics to manage algorithm complexity. Furthermore, there is no generally accepted model of PPIN evolution and unlike with bases in sequence alignment, there is no clear synonymity between proteins.

Most PPIN alignment methods have attempted to align pairs of related proteins, analogously to pairs of similar amino acids in protein sequence alignment. However, many proteins are part of orthologous and paralogous groups. This has been only recently treated in network alignment, due to the significant complications it creates in both the design of an algorithm and in the subsequent assessment of the algorithm's effectiveness. Despite this, a few attempts have been made to create alignment methods that produce many-to-many alignments between proteins; these are exclusively extensions of previous one-to-one alignment methods, such as IsoRankN[86] and Graemlin 2.0[132] (which extend IsoRank[55] and Graemlin[65] respectively). In both of these cases, the later iteration was shown to be more effective, based on functional enrichment of aligned proteins.

2.2 Results

2.2.1 Comparison with PPIN Alignment Algorithms

To assess the GreedyPlus algorithm (see 2.5.1 Algorithm and Figure 2-11 for details), we tested it, along with several algorithms for PPIN alignment, by aligning available worm and yeast SH3 domain IINs.[46,49] We first implemented two naïve alignment algorithms to serve as baselines (see 2.5.2 Comparison Algorithms for details). The first is a greedy algorithm that aligns nodes solely in descending order of similarity score. The second is a seed-and-extend algorithm that initially picks the highest scoring node pair as an initial seed for the alignment. It then extends the

alignment along the edges of the two networks by iteratively aligning the highest scoring pair of unaligned nodes connected to already aligned nodes. We also used several other published network alignment algorithms – IsoRank, GRAAL, H-GRAAL, C-GRAAL, and Natalie 2.0.[147] For fair comparison, the algorithms were prevented from aligning domain nodes and ligand nodes to each other; this was done either using negative scores for domain-ligand pairs or the algorithms were re-implemented with only this specific additional constraint added.

We compare these algorithms' performance based on three metrics. The first two – represented protein orthologies (RPO) and orthologous node pairs (ONP) are measures of how well the algorithms reproduce known orthologous relationships (see Figure 2-1). An RPO is a pair of orthologous proteins, one from each species aligned, which depends on alignment of at least one pair of corresponding interfaces (nodes). An ONP is a pair of aligned interfaces that implies a pair of orthologous proteins; thus, # RPO ≤ # ONP by definition for any alignment. Finally, we ask how well the networks align topologically, by counting the number of edges aligned (EA).

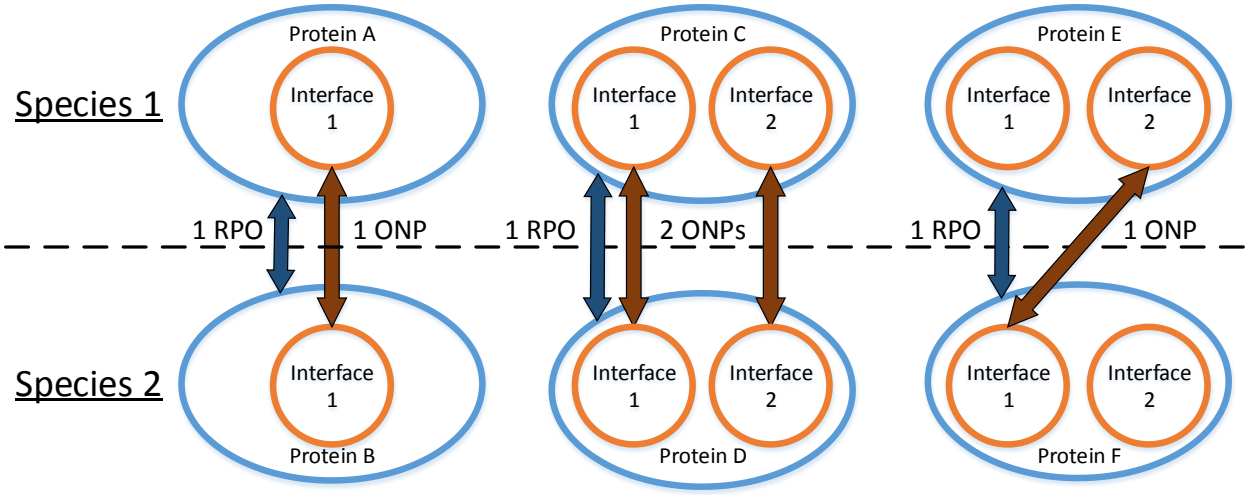


Figure 2-1 – Illustrative examples of represented protein orthologies (RPOs) and orthologous node pairs (ONPs). In each of subfigures, the two proteins are assumed to be orthologous between Species 1 and 2. The orange circles represent specific sites within each protein, depicted as blue ellipses in hypergraph form, and the dark orange arrows represent alignment of the two interfaces. Each pair of aligned interfaces between the two orthologous proteins is 1 ONP. However, regardless of the number of aligned interfaces between the two proteins, there can only be a maximum of 1 RPO, depicted as a dark blue arrow, indicating that the orthologous relationship between the proteins is represented in the alignment.

As IsoRank and Natalie 2.0 use BLAST protein similarity as their only similarity feature, our first comparison uses only BLAST protein similarity. C-GRAAL uses BLAST score among others, so we include it in these tests. The Edge Alignment Weight (EAW) parameter for

GreedyPlus was set to 0.5 after testing with several values (see Discussion). The Greedy, GreedyPlus, and IsoRank algorithms all align similar numbers of orthologous nodes (ONP, 20, 18, and 19 respectively, out of a maximum of 22, see Table 2-1), capturing most of the orthologous protein pairs (RPO) present in our worm and yeast datasets (13, 14, and 12 respectively, out of a maximum of 16).

Alignment using BLAST node similarity	Greedy	Seed & Extend	GreedyPlus	C-GRAAL	IsoRank	Natalie 2.0
# Represented Protein Orthologies (RPO)	13/16 (81%)	1/16 (6%)	**14/16 (88%)**	3/16 (19%)	12/16 (75%)	0/16 (0%)
# Orthologous Node Pairs (ONP)	**20/22 (91%)**	1/22 (5%)	18/22 (82%)	4/22 (18%)	19/22 (86%)	0/22 (0%)
# Edges Aligned (EA)	27/466 (6%)	9/466 (2%)	291/466 (62%)	221/466 (47%)	96/466 (21%)	**354/466 (76%)**

Table 2-1 – Comparison between GreedyPlus, C-GRAAL, IsoRank, and Natalie 2.0 with *C. elegans* and *S. cerevisiae* SH3-mediated IINs. Only BLAST protein scores were used as a similarity feature. The maximum possible values are RPO: 16, ONP: 22, and EA: 466. Bold numbers indicate maximums per row. RPO is the number of known protein orthologies that contain aligned interfaces. ONP is the number of aligned interfaces within orthologous proteins. By definition, ONP ≥ RPO.

While the greedy algorithm was successful at aligning nodes from orthologous proteins, the low (27 out of a maximum possible 466, 6%) number of edges aligned implies that it is a poor network alignment strategy. This may be expected, as the algorithm does not consider edges. The IsoRank algorithm also aligns edges poorly (96 of 466 EA, 21%), as it primarily focuses on the alignment of similar nodes. The bipartite nature of the networks also causes unusual behaviour: the R similarity score in IsoRank fails to properly distribute itself throughout the networks, instead oscillating between domains and ligand sites rather than converging to a stable state. An examination of the resulting IsoRank alignment (see Figure 2-2) reveals no connected concentrations of aligned nodes and edges, and thus no regions of similar network topology between the *C. elegans* and *S. cerevisiae* networks.

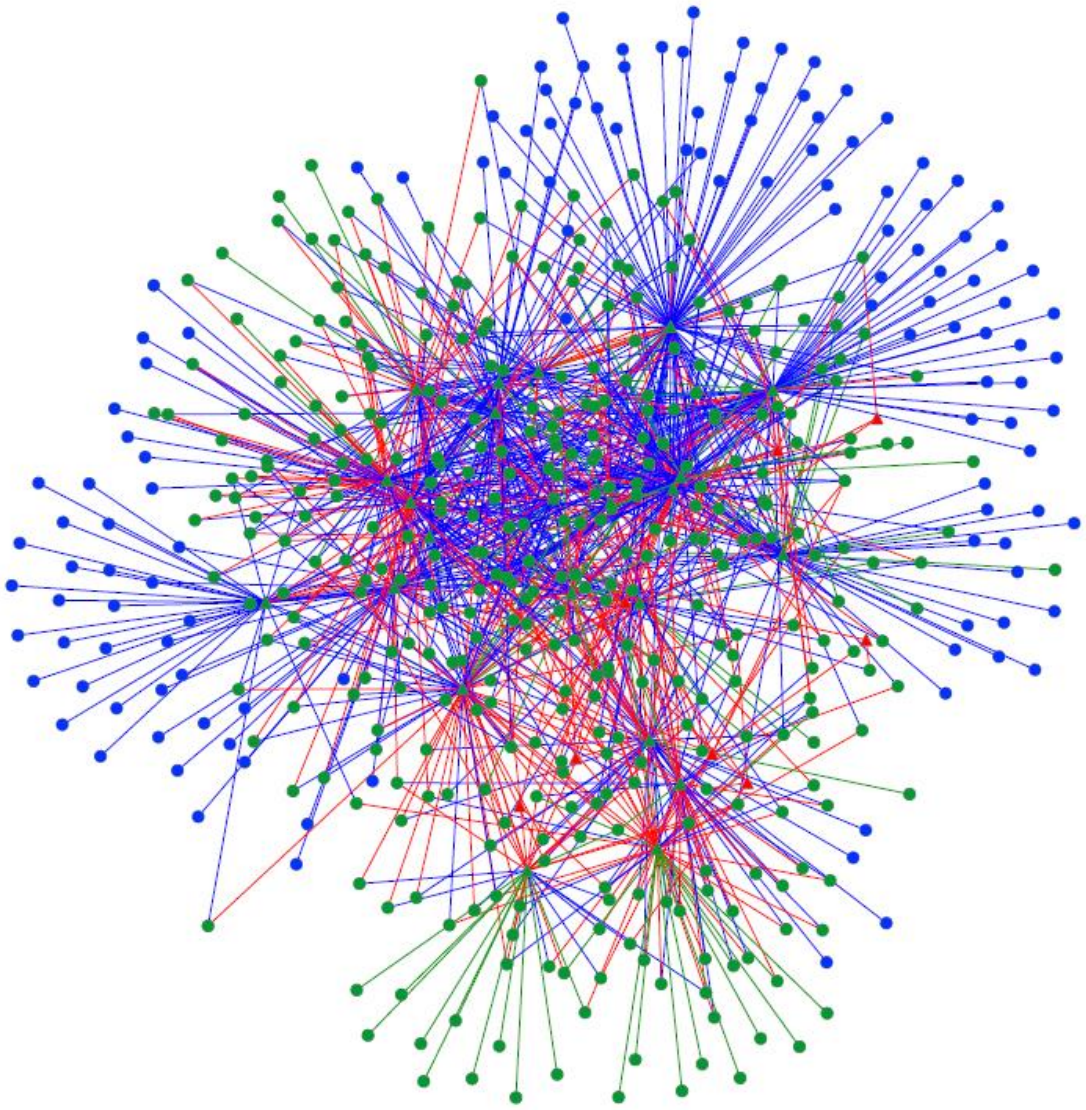


Figure 2-2 – IsoRank alignment of worm and yeast SH3-mediated IINs, using only protein BLAST. Domain interfaces are represented by triangular nodes, ligand interfaces by circular nodes. Yellow nodes are aligned and from orthologous proteins (ONPs), green nodes and edges are aligned but not orthologous, red are unaligned from worm, blue are unaligned from yeast. Node size indicates score. The fact that IsoRank largely ignores edge alignment is reflected in the low number of green edges. While there are more blue edges and nodes than red, due to the larger size of the yeast network, there are no large clusters of green (aligned regions).

Similar to IsoRank, the C-GRAAL algorithm also struggles with the topology of the network. The seeds it finds are invariably ligand sites adjacent to multiple domains, as domains have much higher degrees. However, domains share few common neighbours due to their binding specificity, and ligand sites have generally few neighbours. This limits the alignment expansion, based on its core common-neighbours concept, to less than half of the final alignment.

The seed & extend and Natalie 2.0 algorithms captured very few orthologies (1 and 0 of 16 RPOs, respectively) as they are primarily focused on edge alignment. Seed & Extend makes an early unrecoverable error, beginning alignment at the periphery of the worm network and then rapidly dead-ending, aligning only a total of ten nodes. Natalie 2.0 utilizes a scoring scheme focused on maximizing edge-correctness, on which it performs well at the expense of orthology recovery, indicating that simply maximizing network overlap is insufficient for reproducing known biological relationships.

Finally, GreedyPlus performs best on RPO and second-best on both ONP and EA (14 of 16, 18 of 22, and 291 of 466 respectively). It is also the only algorithm that performs evenly across the three metrics, with performance each at >60% of max, and thus generally performs the best in this comparison (see Figure 2-3).

The GRAAL and H-GRAAL algorithms rely on a single node similarity feature, known as the graphlet degree signature.[70] Thus our second comparison uses only graphlet degree node similarity across all compared algorithms (see Table 2-2), including C-GRAAL as it was also tested with just graphlet degree signature. This node similarity measure results in poor node alignment performance across all algorithms. For example, the GRAAL algorithm identifies no orthologous node pairs, though it does have a similar execution time as GreedyPlus. The generation of the graphlet degree signature for a given node involves counting the number of 2-, 3-, 4-, and 5-node graphlets in which the node participates. However, of the 29 graphlets of such size, 20 of them contain odd cycles not present in bipartite networks. This reduces the number of graphlet orbits, and the length of the graphlet degree signature vector, from 72 to 20. Due to this loss of resolution, the GRAAL algorithm loses power in discriminating between node pairings (see Figure 2-4). Furthermore, the exaggerated spoke-hub network of IINs in comparison to PPINs, for which the GRAAL algorithm was designed, results in the GRAAL algorithm preferring to align non-orthologous nodes to orthologous ones.

H-GRAAL, which focuses on aligning nodes, also fails due to the loss of resolution, while the C-GRAAL algorithm suffers from the same issues as in the earlier tests. Thus, for graphlet degree, different network types require different network topology considerations to be aligned properly.

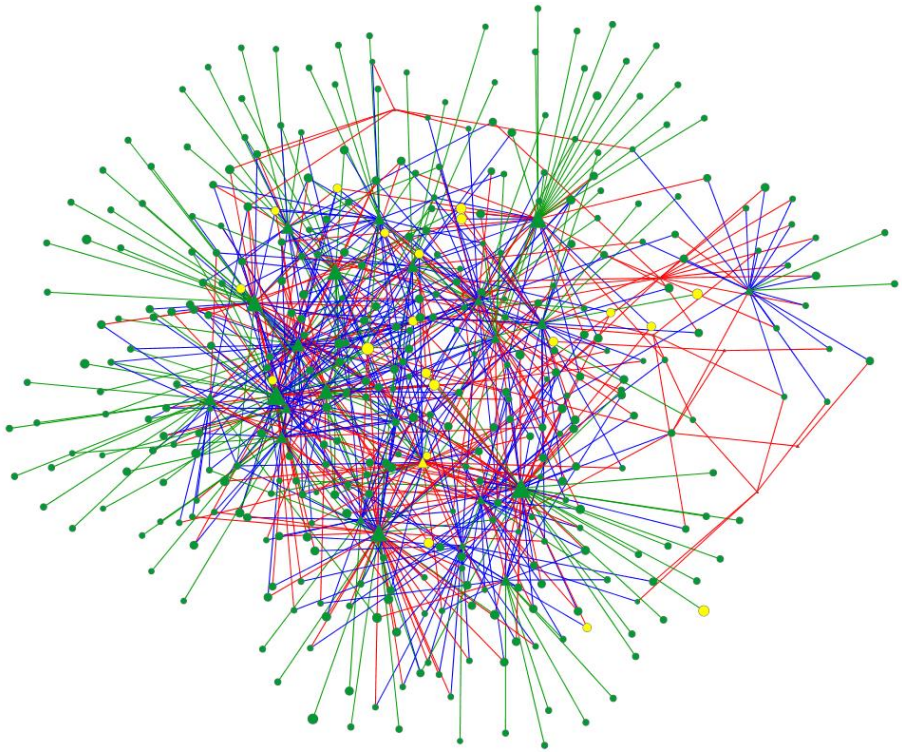


Figure 2-3 – GreedyPlus alignment of worm and yeast SH3-mediated IINs, using only protein BLAST, EAW = 0.5. Domain interfaces are represented by triangular nodes, ligand interfaces by circular nodes. Yellow nodes are aligned and from orthologous proteins (ONPs), green nodes and edges are aligned but not orthologous, red are unaligned from worm, blue are unaligned from yeast. Node size indicates score. The GreedyPlus algorithm aligns many more edges than IsoRank, resulting in many fewer blue and red edges, as they are replaced by half as many green edges. However, there are still no large clusters of green, with red and blue edges dispersed throughout the alignment, indicating that interaction rewiring is both common and distributed.

Alignment using graphlet degree node similarity	Greedy	Seed & Extend	GreedyPlus	C-GRAAL	GRAAL	H-GRAAL
# Represented Protein Orthologies (RPO)	**1/16 (6%)**	0/16 (0%)	0/16 (0%)	0/16 (0%)	0/16 (0%)	**1/16 (6%)**
# Orthologous Node Pairs (ONP)	**1/22 (5%)**	0/22 (0%)	0/22 (0%)	0/22 (0%)	0/22 (0%)	**1/22 (5%)**
# Edges Aligned (EA)	91/466 (20%)	**319/466 (68%)**	298/466 (64%)	295/466 (63%)	157/466 (34%)	93/466 (20%)

Table 2-2 – Comparison between GreedyPlus, C-GRAAL, GRAAL, and H-GRAAL with *C. elegans* and *S. cerevisiae* SH3-mediated IINs. Only graphlet similarity scores were used as a similarity feature. The maximum possible values are RPO: 16, ONP: 22, and EA: 466. Bold numbers indicate maximums per column. RPO is the number of known protein orthologies that contain aligned interfaces. ONP is the number of aligned interfaces within orthologous proteins. By definition, ONP ≥ RPO.

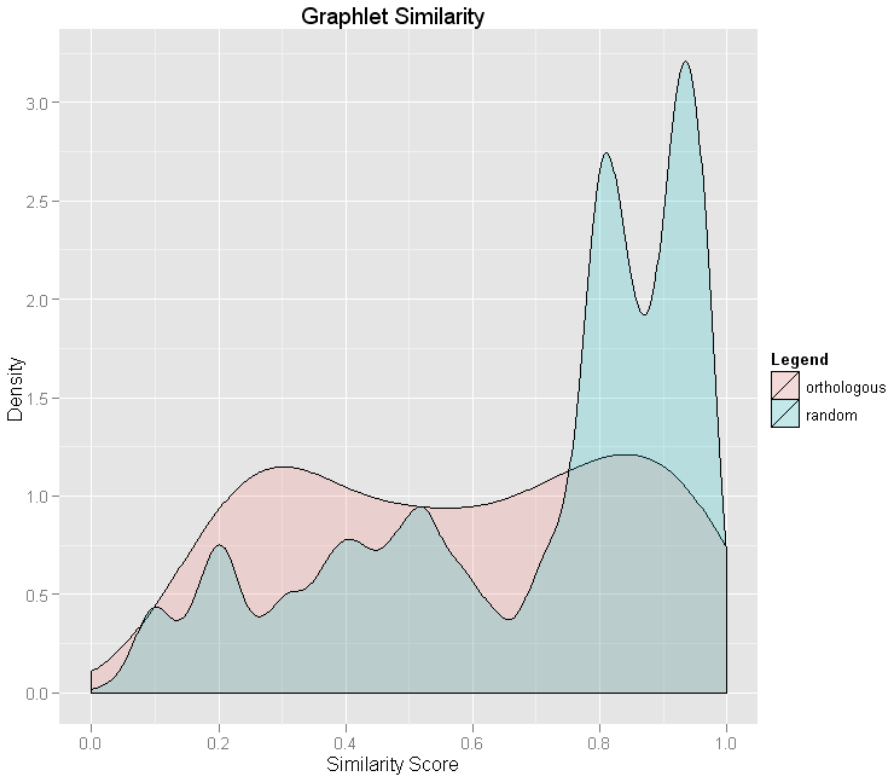


Figure 2-4 – A density plot of graphlet similarity scores between orthologous nodes and random node pairs. Orthologous node pairs (in pink) do not demonstrate a characteristic graphlet similarity score; as such, graphlet similarity has reduced power in correctly aligning nodes. Note that many random node pairs (in blue) have high graphlet similarity in the IINs under study; this is due to the prevalence of leaf nodes, which tend to exhibit similar graphlet degree vectors.

2.2.2 Incorporating More Similarity Features

As seen in our comparisons above, the choice of similarity feature can dramatically affect algorithmic performance across a range of algorithms for a given network. In particular, the simple Greedy algorithm, using a highly informative similarity feature (BLAST), was more successful at recovering orthologous protein relationships than the more advanced GRAAL algorithm using a poor similarity feature (graphlet degree) for our IIN (13 RPOs, 20 ONPs vs. 0 RPOs, 0 ONPs, respectively). To investigate the information content of diverse similarity features and their impact on network alignment, we gathered 29 node similarity features, based on sequence, functional annotation, and network topology characteristics (see Supplementary Table 1, Methods), and attempted alignment with GreedyPlus using all these features, equally weighted with each other and the Edge Alignment Weight (EAW, see 2.5 Methods).

In this comparison, with 29 equally weighted similarity measures, the H-GRAAL algorithm performs best in aligning orthologous nodes, with 11 RPOs and 15 ONPs, but it aligns only 10% (47) of the edges (see Table 2-3). This performance is very similar to that of the Greedy algorithm in all respects, suggesting that edge alignment in H-GRAAL is largely by chance. Other than Natalie 2.0, which produces the exact same alignment as in the first comparison (see Table 2-1), the Seed & Extend algorithm aligns the fewest orthologous nodes (2 ONPs, 9%), but aligns the most edges (306 EAs, 65%). As every pair of nodes the algorithm aligns must be connected to two previously aligned nodes, the algorithm tends to generate a high number of edge alignments, but this inflexibility causes it to ignore possible node alignments supported by high similarity scores when no neighbours have yet been aligned. The importance of the input similarity feature is shown by the improved performance of GRAAL due to the introduction of more informative similarity features, and the decreased performance of IsoRank, due to the dilution of the highly informative BLAST similarity feature. However, GreedyPlus had the best balanced performance in both properly aligning node pairs (44% RPOs, 45% ONPs) and the number of aligned edges (51% EAs).

Alignment using 29 equal weight node similarity measures	Greedy	Seed & Extend	Greedy-Plus	C-GRAAL	GRAAL	H-GRAAL	Iso-Rank	Natalie 2.0
# Represented Protein Orthologies (RPO)	10/16 (63%)	2/16 (13%)	7/16 (44%)	1/16 (6%)	4/16 (25%)	**11/16 (69%)**	7/16 (44%)	0/16 (0%)
# Orthologous Node Pairs (ONP)	13/22 (59%)	2/22 (9%)	10/22 (45%)	1/22 (5%)	5/22 (23%)	**15/22 (68%)**	9/22 (41%)	0/22 (0%)
# Edges Aligned (EA)	35/466 (8%)	305/466 (65%)	238/466 (51%)	293/466 (63%)	56/466 (12%)	47/466 (10%)	87/466 (19%)	**354/466 (76%)**
Runtime (ms)	766	794	2,719	2,755	**2,695**	84,722	112,289	1,804,620*

Table 2-3 – Alignment algorithm performance on *C. elegans* and *S. cerevisiae* SH3-mediated IINs using all similarity features. All 29 similarity features were used with naïve parameterization. The maximum possible values are RPO: 16, ONP: 22, and EA: 466. Bold numbers indicate maximums per row (Greedy and Seed & Extend are excluded from runtime comparison). RPO is the number of known protein orthologies that contain aligned interfaces. ONP is the number of aligned interfaces within orthologous proteins. By definition, ONP ≥ RPO. * The original distribution was used for Natalie 2.0. All other algorithms were implemented in Java by the authors.

2.2.3 Parameter Weight Tuning

Having established GreedyPlus' performance using naïve parameterizations on the weights of each of the 29 similarity measures, we investigated how improved parameterizations would affect alignment quality. We used a random-restart hill-climbing strategy to search the high-dimensional parameter space for local maxima in orthology recovery (see 2.5 Methods). This strategy was applied to all 29 similarity features plus the edge alignment weight together (see Table 2-4). Using this procedure, we found a set of parameters that can recover all possible orthologies (16 RPO, 21 ONP) with a high number of edges aligned (210/466, 45%).

Proteins			
BLAST coverage	8.48	BLAST score	6.89
TCSS - biological process	0.62	TCSS - cellular component	2.56
TCSS - molecular function	6.98		
Domains			
Average shortest path length	7.60	Betweenness centrality	5.12
BLAST coverage	6.71	BLAST score	2.65
Closeness centrality	6.54	Degree	0.53
Eccentricity	1.24	Graphlet degree similarity	7.16
Neighbourhood connectivity	2.74	Radiality	4.42
Stress	7.69	Topological coefficient	1.86
Ligands			
Average shortest path length	1.94	Betweenness centrality	4.06
Closeness centrality	0.88	Degree	0.17
Eccentricity	1.50	Graphlet degree similarity	0.27
Neighbourhood connectivity	0.53	Radiality	0.44
Smith-Waterman coverage	5.04	Smith-Waterman score	1.33
Stress	1.50	Topological coefficient	0.71
Edge alignment weight	1.86		

Table 2-4 – An "optimal" parameter set for GreedyPlus, normalized out of 100. The full set of similarity features tested with GreedyPlus, and the corresponding weights used to achieve "optimal" alignment performance.

However, several local maxima existed that each resulted in similarly high orthology recovery. Also, some parameters are similar to each other, thus not all 29 may be required. To address the possibility of overfitting, we gradually reduced the number of parameters while repeating the search/optimization procedure. In so doing, we found a set of 6 parameters still capable of

producing high-quality alignments (16 RPO, 22 ONP, 218/466 or 47% EA), as shown in Table 2-5.

Proteins			
BLAST score	46.86	TCSS – molecular function	5.71
Domains			
BLAST score	1.14	Closeness centrality	20.71
Ligands			
Closeness centrality	1.71	Smith-Waterman score	16.00
Edge alignment weight	8.00		

Table 2-5 – A reduced "optimal" parameter set for GreedyPlus, normalized out of 100. A reduced set of similarity features used by GreedyPlus to achieve "optimal" alignment performance, and their associated weights.

Sequence similarity features account for ~64% of the overall parameter weighting. Network topology considerations, including the closeness similarity features and the edge alignment weight – which is not a similarity feature and can be applied multiple times to the same pair of potentially aligned nodes – account for ~30%.

Including closeness and the edge alignment weight in addition to the sequence similarity features improved orthology recovery. When closeness was removed, the resulting alignment produces only 13 RPO, 17 ONP, and 231 EA. Similarly, setting the edge alignment weight to zero results in a poorer alignment, in particular with edges: 13 RPO, 19 ONP, 34 EA. The small weight assigned to the functional similarity feature Topological Clustering Semantic Similarity (TCSS),[148] however, is insignificant; setting it to zero did not change the overall alignment performance, despite TCSS being weighted relatively highly when optimization was performed using all assembled features (see Table 2-4). Removing all non-sequence similarity features (i.e. using only BLAST and Smith-Waterman) results in 13 RPO, 19 ONP, 29 EA. Given the decreased performance using just sequence similarity features, we conclude that non-sequence similarity features are useful in determining the similarity between nodes for the purposes of network alignment.

2.2.4 A Zoom-in

As a snapshot of how GreedyPlus works in practice, we zoom in on the yeast protein BZZ1 (see Figure 2-5), which has two domain nodes in our dataset. BZZ1 is a recruiter protein involved in

regulating actin polymerization,[149] and is an ortholog of the worm protein SDPN-1; Gene Ontology identifies both genes as involved in endocytosis[92]. In this alignment, performed by GreedyPlus using its tuned similarity weights, one of the BZZ1 domains is aligned to SDPN-1's single SH3 domain. However, because GreedyPlus cannot perform one-to-many alignments, BZZ1's other SH3 domain is aligned to EPHX-1, which is not an identified ortholog. Neither SDPN-1 nor EPHX-1 are among BZZ1's top BLAST scores, ranking 9th and 11th among our dataset; however, the other similarity features and the Edge Alignment Weight drive up their priority in alignment.

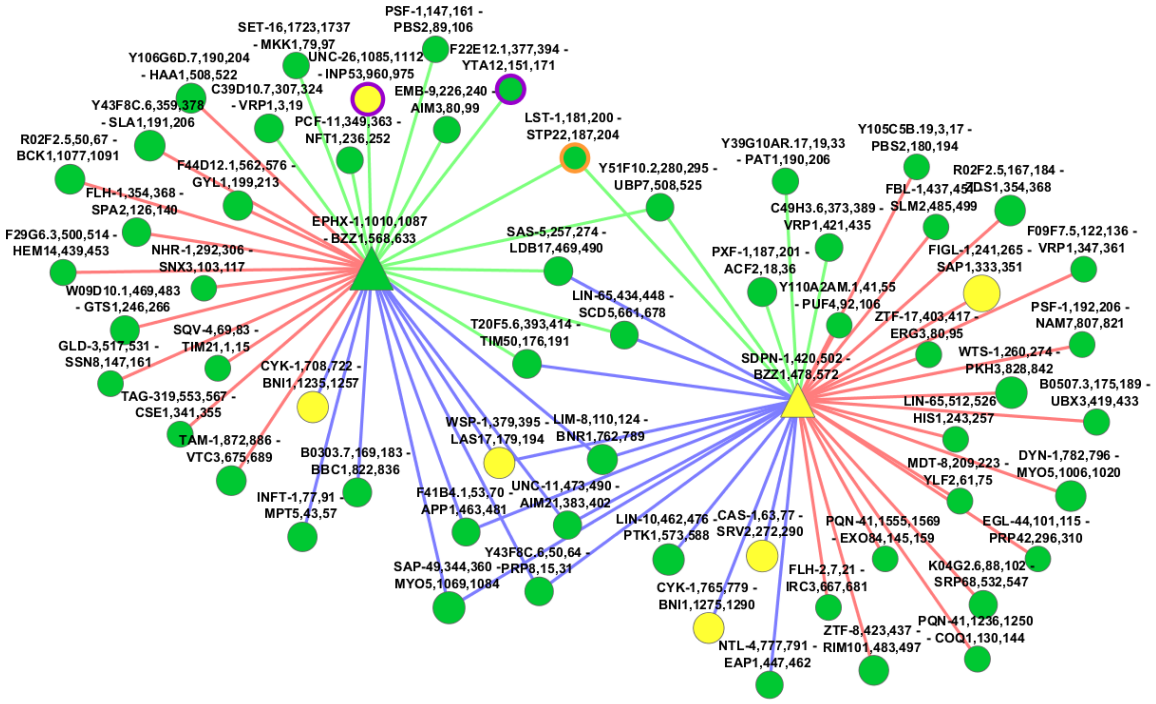


Figure 2-5 – A zoom-in of the "optimal" GreedyPlus alignment of worm and yeast SH3-mediated IINs, consisting of the two yeast BZZ1 nodes and all their neighbours. Domain interfaces are represented by triangular nodes, ligand interfaces by circular nodes. Yellow nodes are aligned and orthologous; green nodes and edges are aligned, red are unaligned from worm, blue are unaligned from yeast. Node size indicates score. The two EAW contributors to EPHX-1,1010,1087 - BZZ1,568,633 are outlined in purple; the EAW contributor to SDPN-1,420,502 - BZZ1,478,572 is outlined in orange.

Interestingly, the EPHX-1 – BZZ1 alignment was performed first, as the neighbouring aligned pairs F22E12.1,377,394 - YTA12,151,171 and UNC-26,1085,1112 - INP53,960,975 boost its score, via the EAW, by ~36%, illustrating the additive effect of the EAW. Subsequently, the nodes LST-1,181,200 and STP22,187,204 are aligned, partially on the strength of the EAW from

EPHX-1 – BZZ1, which then promotes the alignment of SDPN-1 to BZZ1. A number of other orthologous node alignments occur in the immediate neighbourhood, but do not contribute any EAW to the BZZ1 domain alignments because they are not adjacent in either the worm or yeast networks. For example, WSP-1 and LAS17 are orthologs, but while LAS-17 interacts with BZZ1 in yeast, its worm ortholog WSP-1 does not interact with either SDPN-1 or EPHX-1 in our SH3 dataset, nor is such an interaction found in the interaction data iRefIndex,[150] hinting at a previously undetected interaction.

We also observe that while BNI1 is an interaction partner with BZZ1, with two sites targeted by the two BZZ1 SH3 domains, its worm ortholog CYK-1 does not interact with either EPHX-1 or SDPN-1. This non-interaction is also supported by iRefIndex. In our worm network, the respective CYK-1 sites are targeted only by Y106G6H.14 and TOCA-1, neither of which have functional annotations in GO, though TOCA-1 is indicated to be involved in endocytosis as well[151,152] (see Figure 2-6). This extensive interaction rewiring suggests that IIN alignment approaches based on maximizing network topology overlap may not be appropriate in identifying orthologs.

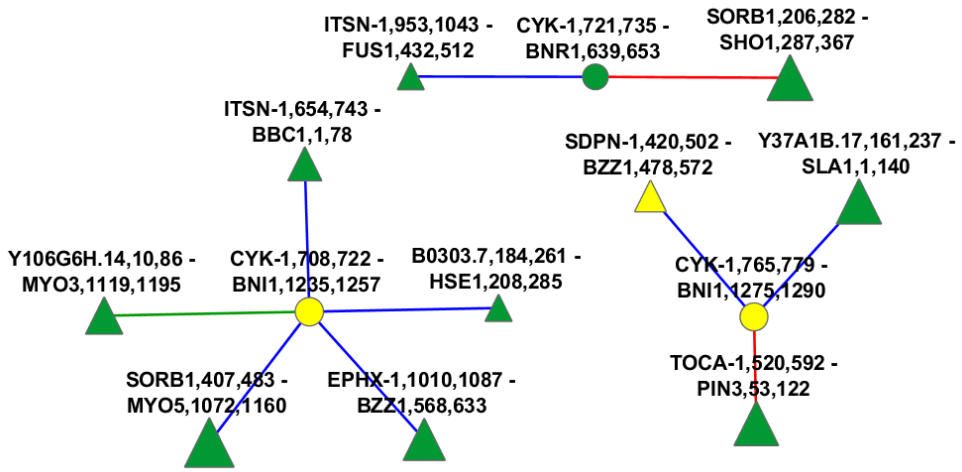


Figure 2-6 – A zoom-in of the "optimal" GreedyPlus alignment of worm and yeast SH3-mediated IINs, consisting of the three worm CYK1 nodes and all their neighbours. Domain interfaces are represented by triangular nodes, ligand interfaces by circular nodes. Yellow nodes are aligned and orthologous; green nodes and edges are aligned, red are unaligned from worm, blue are unaligned from yeast. Node size indicates score.

2.2.5 Yeast Subspecies Alignments

In addition to the *C. elegans* to *S. cerevisiae* IIN alignment, we tested GreedyPlus on published predicted SH3 IINs from 18 different yeast species[105] (see 2.5. Methods). All 18 networks were pairwise aligned, using both the full set of features and the reduced set identified above. TCSS was removed as a similarity feature to remove circularity, as most GO annotations for these yeast species' proteins are predicted via orthology with *S. cerevisiae*. The feature weighting identified in the above described optimization for GreedyPlus on *C. elegans* and *S. cerevisiae* was used.

As these species are more closely related than *C. elegans* and *S. cerevisiae*, we found, as expected, that GreedyPlus is able to recover more orthologous pairs in these pairwise alignments. When using a minimal set of similarity features with optimized weights (see Table 2-5), GreedyPlus alignments almost always recovered more than 70% of the known orthologous protein pairs while still maintaining a high percentage of edges aligned (mean 50.6% of maximum possible, see Figure 2-7). Using all the gathered similarity features, except TCSS (26 features), GreedyPlus still performed well, aligning an average of 56% of orthologous protein pairs (see Figure 2-8).

In both cases, a high percentage of the edges were aligned; notably, more edges were aligned when the full set of 26 similarity features were considered. This result is contrary to what would be expected; given equal weights, with more similarity features, the relative weight of the EAW is decreased from $1/5^{th}$ of the overall scoring function to $1/26^{th}$. This suggests that the additional similarity features – almost all based on network topology – increase the alignment of edges by promoting the alignment of nodes with similar local network topology. This may be due to the inferred nature of these networks from relatively closely related species which lead to networks with unusually similar network topologies.

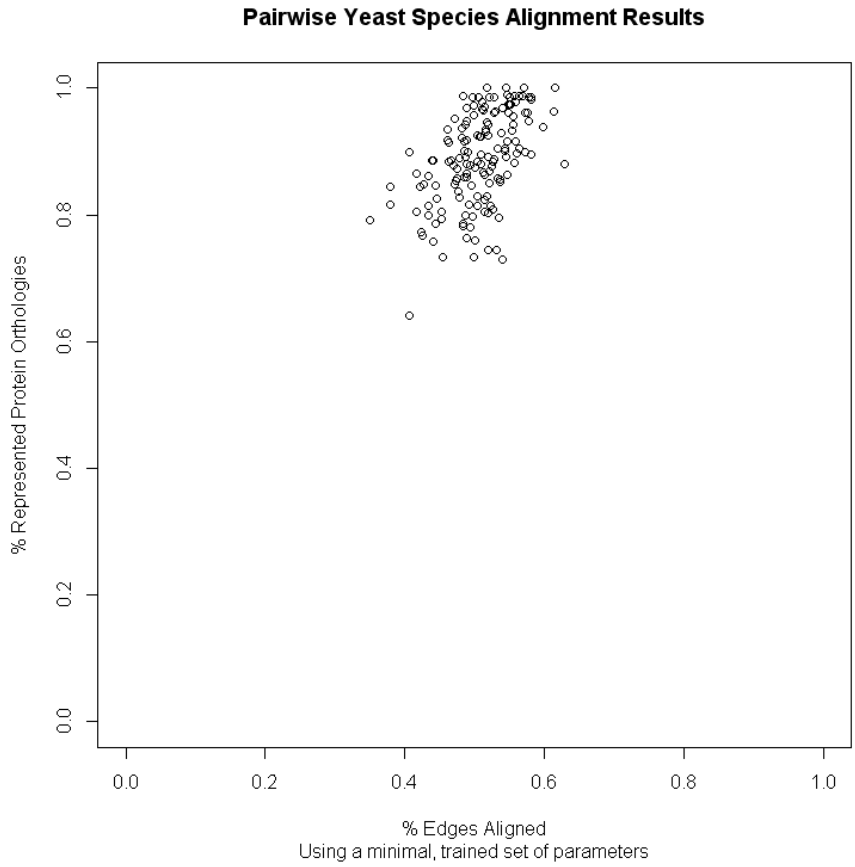


Figure 2-7 – Percent RPO and EA achieved for pairwise yeast species alignments. Using the optimized parameters from Table **2-5**, GreedyPlus was run on each pair of yeast networks (see 2.5 Methods). The percent of represented protein orthologies and edges aligned for each alignment was retrieved and plotted on the same scale.

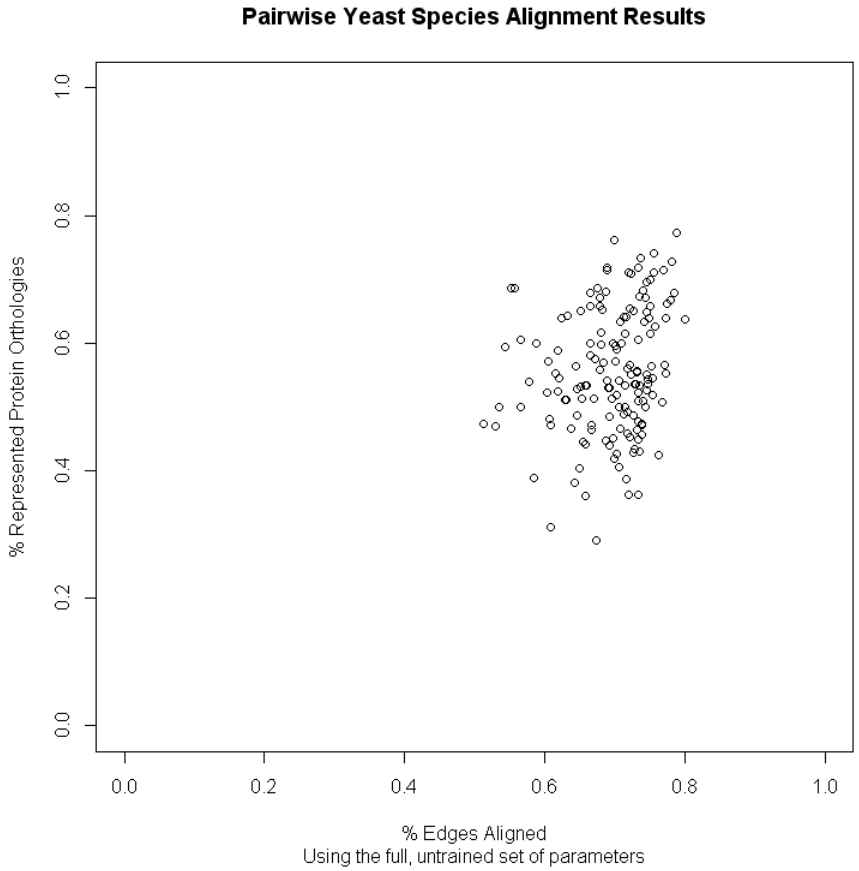


Figure 2-8 – Percent RPO and EA achieved for pairwise yeast species alignments. Using the full set of similarity features and no optimization, GreedyPlus was run on each pair of yeast networks (see Methods). The percent of represented protein orthologies and edges aligned for each alignment was retrieved and plotted on the same scale.

2.3 Discussion

We have described GreedyPlus, a network alignment algorithm that is effective, flexible in terms of input data, and fast, outperforming traditional network alignment methods in aligning IINs. With feature optimization, made easier by GreedyPlus' speed, we identified a set of data features and their weights that proved highly effective in guiding network alignment.

Unlike other network alignment algorithms, GreedyPlus explicitly specifies a trade-off between a node alignment and edge alignment via the EAW parameter. This means that *a priori* knowledge about the networks being aligned can be used to control alignment. Lower EAW values should be more suitable for dissimilar networks, to force the algorithm to focus on nodes, whose similarity scores should already be sufficiently differentiated to distinguish proper from improper

alignments. On the other hand, higher EAW values should be more suitable for highly similar networks, resulting in a stricter alignment, which would highlight the few areas of difference.

This feature makes the GreedyPlus algorithm suited for evaluating the relative importance of node versus edge alignment in network alignments. Identifying the correct parameterisation for the alignment of different types of networks is in itself an interesting research problem capable of informing us on how networks evolve.

Another important feature of GreedyPlus is that it is mostly agnostic to the topological nature of the networks being aligned, other than the assumption that neighbours of aligned nodes should more likely be aligned themselves. As our SH3 domain data set does not contain domain-domain or ligand-ligand interactions, the IINs we studied were bipartite, which confounded several of the algorithms tested. Though the current GreedyPlus implementation is specialized to handle bipartite networks, it is not dependent on the bipartition, and the approach could be adapted to different network types. For example, domain-domain interactions are possible with other domains, such as SAM and coiled-coil, and so IINs are not necessarily bipartite.

Though we lack sufficient IIN data to make a general statement, we observed a trade-off between the alignment of biologically verified node pairs and the alignment of edges with a number of algorithms, including our own. Notably, an increase in the number of edges aligned did not necessarily lead to an increase in the number of nodes properly aligned. Some brief experiments with the Edge Alignment Weight parameter showed that, with GreedyPlus, attempting to maximize the number of aligned edges results in a distinct decrease in the number of properly aligned nodes (see Figure 2-9).

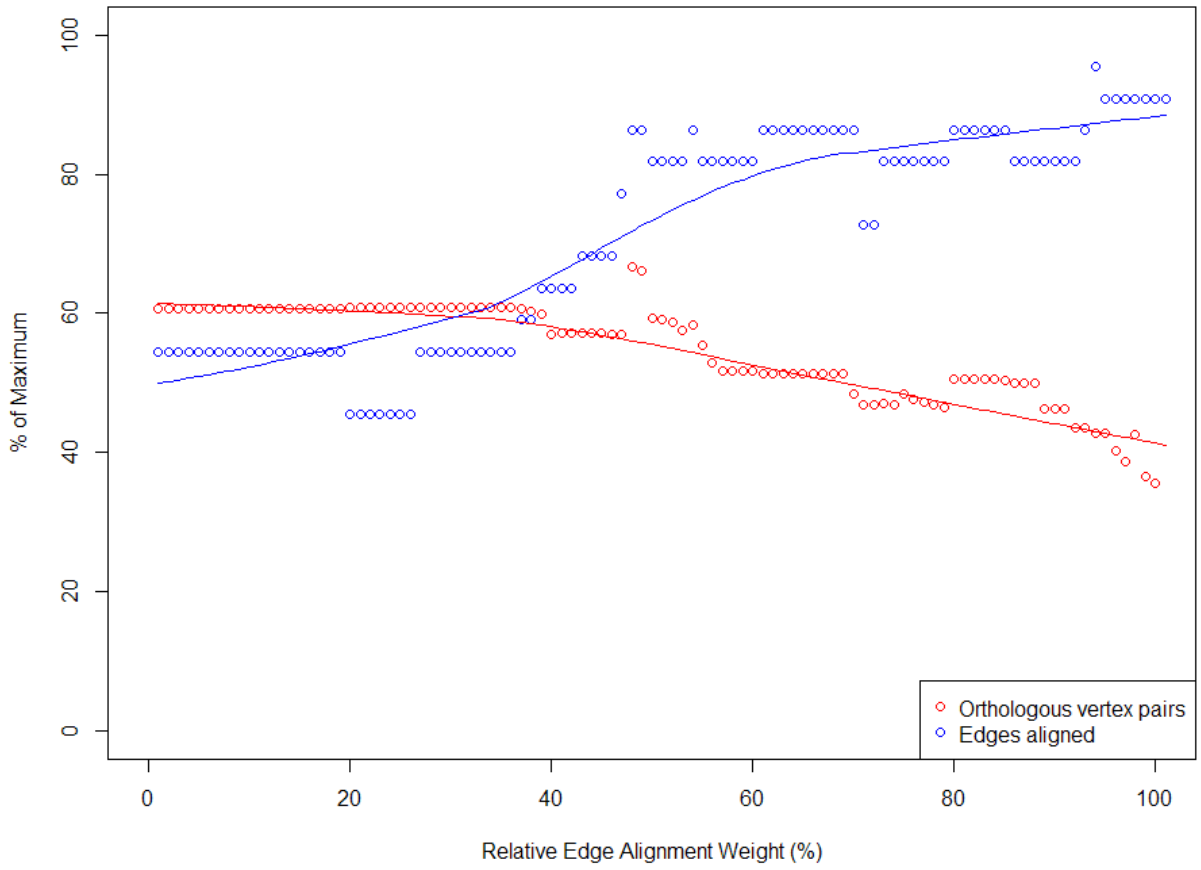


Figure 2-9 – Trade-off in GreedyPlus performance between orthologous nodes aligned and edges aligned. Using BLAST score between proteins as the only similarity feature, we ran GreedyPlus with the edge alignment weight set at values ranging from 0% to 100% of the protein BLAST score weight, and plotted its performance. If the plots were tightly correlated, it would indicate that successfully aligning networks by network topology would be equivalent to aligning node pairs successfully. However, we observe a distinct trade-off between aligning edges and aligning orthologous nodes.

Despite this trade-off, our results show that including network topology similarity features improves the orthology predictions of network alignment, demonstrating their relevance, and hinting that network alignment may complement sequence alignment as a bioinformatics tool to study evolution, but the significance of edge alignment in constructing network alignments is unclear. While aligning two edges implies similarity between their endpoints, simply maximizing the number of edges aligned clearly does not result in a biologically relevant and informative alignment. Though it would be a simple extension, GreedyPlus does not currently implement edge similarity features, and treats all interactions as functionally identical, because while some are available, such as PPI confidence estimates,[153] binding affinity, tissue specificity, or the types of interacting residues, they are not currently prolific enough to be generally useful.

It is possible that with discriminatory information about the interactions that edges represent, a method based on optimizing edge alignments may prove effective for aligning entire networks as well. Using this information, a PPIN alignment algorithm, for example, could preferentially align two SH3-mediated interactions, rather than an SH3-mediated interaction and a WW-mediated interaction, or could avoid aligning a structural interaction to a transient enzymatic interaction.

We selected a broad set of similarity features to investigate their utility in network alignment. Sequence similarity, in particular BLAST-based, is ubiquitous in network alignment research; ligand sequences, however, are too short for BLAST, so Smith-Waterman was used instead. GO functional annotations are often used to validate network alignments; we were interested in examining whether they could be used as an input feature as well. Network topology features other than graphlet signatures have had some treatment in the literature; Kuchaiev et al.[72] have previously evaluated graphlet signature, node degree, clustering coefficient, and eccentricity. As such, we opted to incorporate many different network topology features to assess their utility (see Table 2-4). In general, the features are of two types: centrality or clustering. Average shortest path length, betweenness, closeness, eccentricity, radiality, and stress are all measures whether a node is located in the centre or on the periphery of a network. The alignment of a central node to a peripheral node would, given our current knowledge of network evolution, imply a highly improbable evolutionary history, involving a large redistribution of node and edges about a previously peripheral node. Degree, graphlet signature, neighbourhood connectivity, stress, and topological coefficient are all measures of connectedness in the neighbourhood around a node. These measures should distinguish a node in a highly connected neighbourhood from one in a sparse neighbourhood, providing regional information to guide alignment. Clustering coefficient was not used, as the nodes in a bipartite graph always have a clustering coefficient of zero.[123]

Notably, using BLAST score for full-length proteins alone as the single distinguishing similarity feature results in substantial orthology recovery in the alignment (see Table 2-6). Conversely, using the sequence similarity scores of the nodes - BLAST for domains and Smith-Waterman for ligands – alone resulted in no orthology recovery. Various combinations of the network topology and functional similarity features, as determined using TCSS,[148] also resulted in low orthology reproduction. This seems to suggest that while non-sequence-based information has a role to play in network alignment, its direct contribution is not obvious.

	Represented Protein Orthologies	Orthologous Node Pairs	Edges Aligned
Protein BLAST	10	12	299
Domain BLAST / Ligand S-W	0	0	296
Protein & Domain BLAST / Ligand S-W	6	10	288
Network Topology Features	0	0	272
Functional Features	5	7	278

Table 2-6 – GreedyPlus performance using different similarity features. GreedyPlus was used to align the *C. elegans* and *S. cerevisiae* SH3 IINs, using different sets of similarity features under simple parameterization (all scoring weights, including EAW, are equal), and the results of each alignment are shown.

Many measures of network alignment quality are also dependent on BLAST similarity. In addition to orthologs, some measure of coherence between the GO terms of aligned nodes is often used to verify the biological quality of an alignment. However, many GO terms are inferred from another protein based on sequence similarity, either directly or indirectly. Even experimentally derived GO terms may be subject to BLAST-derived confirmation bias, as experimental design could be guided by BLAST results. Notably, while using functional features alone generates an alignment with orthology reproduction (see Table 2-6), they played close to no role in our optimized parameter set (see Table 2-5); this may be explained by a duplication of information between sequence- and function-based similarity features. If network alignment is to serve as an independent tool alongside BLAST, the development of assessment measures that do not involve BLAST-based confirmation bias will be essential.

Finally, network alignments, consisting of at least two networks plus the alignment between them, present a challenging visualization problem.[138] Even with the relatively small, sparse IINs, alignments produce networks exhibiting the same "hairball" nature characteristic of the PPIN visualization problem.[138] For instance, it would be useful to evaluate compound graph and hypergraph visualizations, similar to what is presented in Figure 1. With the proliferation of network alignment research, improved visualization tools will be critical for the interpretation of generated alignments.

2.4 Conclusions

GreedyPlus is a novel algorithm useful for the alignment of interface-interaction networks, compatible with a range of node similarity measures. While node sequence information is dominant in its ability to align nodes properly, network topology information is useful for improving alignment performance, even if it is of low utility in isolation. We identify a reduced set of information types and a weighting of these types that can be used to generate relatively high performance alignments. The algorithm and our evaluation framework will be used to further investigate network evolution and how to best align biological networks.

2.5 Methods

2.5.1 Algorithm

We created a fast algorithm for the alignment of IINs, GreedyPlus, using a local network alignment approach. Additionally, alignments generated by GreedyPlus can easily be compared, on a stepwise basis, to identify when and why each alignment tuple was formed (or not), allowing us to specifically query how changes in parameterization and input data may lead to differences in the resulting alignment. Further, a key research question for network alignment is how to balance node-specific information versus network topology information. The critical characteristic of network alignment is that edges must be aligned in addition to nodes; an alignment that has no aligned edges is, fundamentally, not a network alignment.[154] However, while there is a plethora of biological information regarding proteins, there is a dearth of information on protein-protein interactions that would assist in guiding or verifying an alignment. Thus, to investigate the interrelation and relative importance of node versus edge alignment, GreedyPlus explicitly models a balance between these two elements.

An interface interaction network can be modelled as an undirected graph G, consisting of a node set V and an edge set E, where each edge is a tuple of two nodes (v_1, v_2). An alignment of two PPINs G_1 and G_2 is thus a set of 2-tuples $A = \{[u_1, v_1], [u_2, v_2], \ldots, [u_i, v_i]\}$, where $u_i \in V_1 \in G_1$ and $v_i \in V_2 \in G_2$, if the alignment is one-to-one. In this context, the goal of GreedyPlus, like most other biological network alignment algorithms, is primarily to align nodes $[u_i, v_i]$ such that biological inferences can be drawn about one node from the other based on their alignment. If

these nodes represent interfaces, such an inference might be that they are orthologous, or that they mediate functionally similar interactions, or that they evolved to occupy similar positions in their respective networks under similar selective pressures.

Though generalizable to other networks, GreedyPlus' current implementation is tailored specifically to accommodate bipartite peptide-recognition module-mediated IINs, reflecting the current availability of IIN data. These networks, being bipartite, can be modelled in a similar manner as general IINs, as an undirected graph G, with a node set $V = \{D, L\}$, where D and L are the sets of nodes representing peptide recognition domains (e.g. SH3) and ligands respectively, and $\forall\, v \in G$, $v \in D$ or $v \in L$. An alignment A of such networks is then restricted such that for each tuple $[u, v] \in A$, either $(u \in D \wedge v \in D)$ or $(u \in L \wedge v \in L)$.

The key intuition behind the GreedyPlus algorithm is that the presence of interaction is itself a biological evidence source pointing towards an orthologous relationship between a pair of proteins. That is, if there exists $(u_1, u_2) \in G_1$ and $(v_1, v_2) \in G_2$, and it is known that u_2 and v_2 are orthologous, then we can infer that u_1 and v_1 are more likely to also be orthologous. Furthermore, if there also exist $(u_1, u_3) \in G_1$ and $(v_1, v_3) \in G_2$, this would provide even stronger evidence; thus the more edges that would be aligned by aligning nodes u_1 and v_1, the more likely that this is a good alignment of nodes.

The GreedyPlus algorithm is essentially a greedy algorithm that iteratively aligns pairs of nodes in descending order of similarity, defined by a given similarity score. However, when aligning two nodes, GreedyPlus also considers the number of edges that would be aligned if the nodes were aligned, strengthening the respective node pair similarity score with more edges aligned (see Figure 2-10 and Figure 2-11). Thus, GreedyPlus will prefer to align node pairs that also align edge pairs over those that do not if the difference in similarity is small, but will align highly similar nodes irrespective of network topology. The preference of the algorithm in aligning edges or maximizing node similarity can be controlled using a defined parameter, named the *edge alignment weight (EAW)*, providing flexibility and enabling investigation of the relative importance of aligning nodes versus edges.

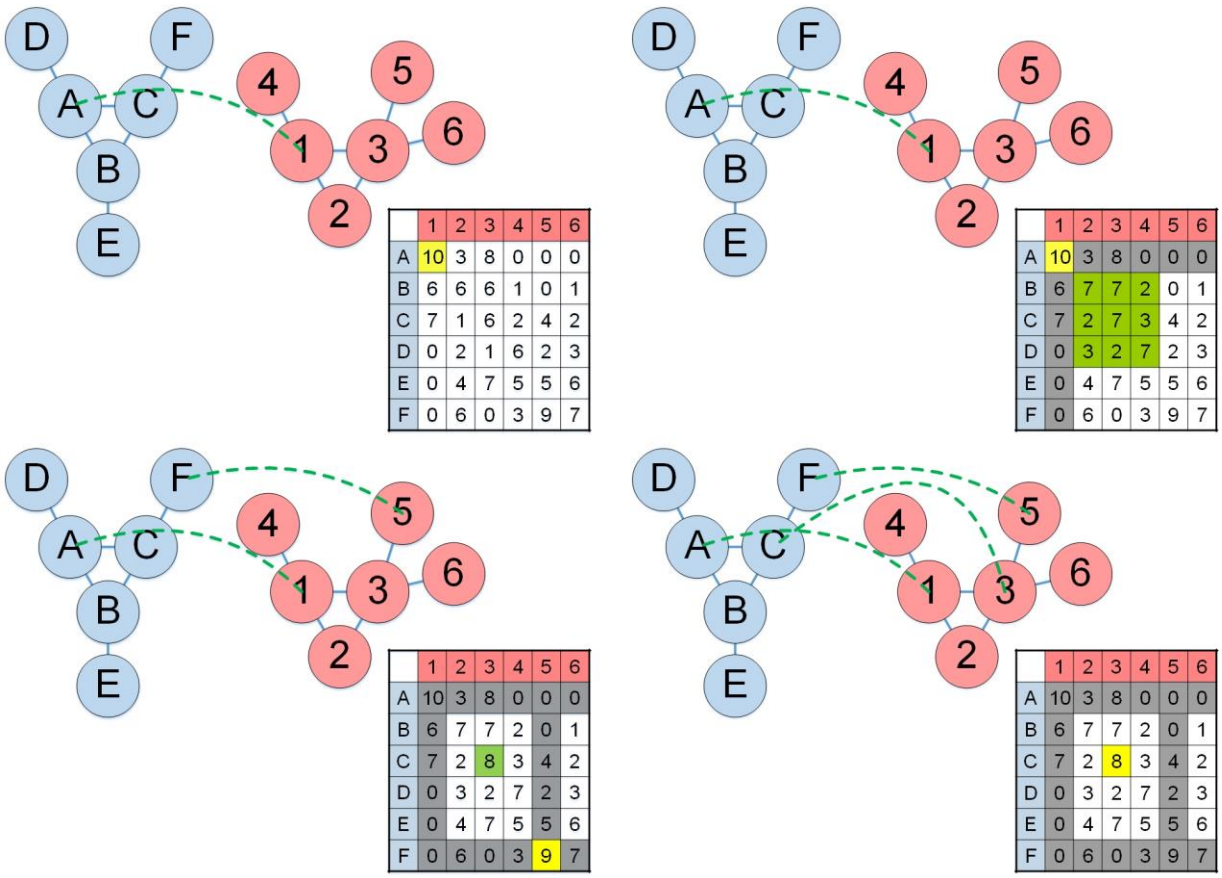


Figure 2-10 – A simple example of GreedyPlus in action. First, GreedyPlus finds the highest scoring pair of nodes (in yellow), in this case the pair (*A, 1)*, and aligns them. Then the similarity matrix is updated, with the scores of all pairs of all neighbours of just-aligned nodes (in green) incremented by the Edge Alignment Weight (in this case, 1). Using the updated similarity matrix, GreedyPlus iterates until all nodes are aligned. In this example, the third node alignment [*C*, *3*] is made as a result of the Edge Alignment Weight increasing its similarity score; otherwise, the pairing [*E*, *3*] would have been made instead.

The edge alignment weight (EAW) determines how strongly GreedyPlus prioritizes the alignment of edges compared to nodes. When the EAW is set to zero, GreedyPlus behaves identically to the greedy alignment algorithm, as it will ignore the alignment of edges. When the EAW is set to ∞, GreedyPlus behaves similarly to a seed-and-extend algorithm (see below), always choosing to align two edges whenever possible as the EAW will overwhelm any preference in aligning nodes, with the exception that it can resume alignment even if edge extension possibilities are exhausted. By tuning the EAW parameter to intermediate values, the preference between node or edge alignment can be balanced.

GreedyPlus Algorithm

```
1: alignment ← null
2: EAWeight ← user input
3: while ∃u, v s.t. [(u ∈ D1 and v ∈ D2) or (u ∈ B1 and v ∈ B2)] and u, v ∉ alignment do
4:    bestPair ← null
5:    for all u, v do
6:       if Score(u, v) + EA(u, v) > Score(bestPair) + EA(bestPair) then
7:          bestPair ← (u, v)
8:       end if
9:       alignment ← alignment + bestPair
10:   end for
11: end while
12:
13: EA(u,v) {
14: EAScore ← 0
15: for all ⟨w, x⟩ ∈ alignment do
16:    if (u, w) ∈ E1 and (v, x) ∈ E2 then
17:       EAScore ← EAScore + EAWeight
18:    end if
19: end for
20: return EAScore
21: }
```

Figure 2-11 – Pseudocode for the GreedyPlus algorithm. This pseudocode is not optimized, for clarity purposes. Note that EAW has an additive effect; the more edge alignments that support a given node alignment, the higher the corresponding score is boosted.

A naïve implementation of GreedyPlus runs in worst-case $O(|D|^3|E| + |L|^3|E|)$ time, though given the nature of domains and ligands, $|L| >> |D|$. In practice, GreedyPlus takes approximately two seconds to align two networks with $|L| = 500$ and $|E| = 600$, implemented in Java 7 on a 3.4 GHz processor.

2.5.2 Comparison Algorithms

To compare with GreedyPlus, we created two simple algorithms for IIN alignment to serve as baselines. The first we deemed the Greedy algorithm (see Figure 2-12). The Greedy algorithm simply aligns the highest scoring pair of nodes between the two networks repeatedly and exhaustively. It pays no attention to the alignment of edges, and the only graph theoretic considerations in its alignment are those embedded within the scoring function used. Effectively, the Greedy algorithm performs identically to the GreedyPlus algorithm, when the latter's edge alignment weight is set to 0.

Greedy Algorithm

```
alignment ← null
while ∃u, v s.t. u ∈ D1 ∪ B1 and u ∉ alignment and v ∈ D2 ∪ B2 and v ∉ alignment do
    bestPair ← null
    for all u, v do
        if Score(u, v) > Score(bestPair) then
            bestPair ← (u, v)
        end if
        alignment ← alignment + bestPair
    end for
end while
```

Figure 2-12 – The Greedy algorithm for IIN alignment. This is a simple algorithm for IIN alignment used as a comparison baseline for GreedyPlus. It aligns the highest scoring pair of nodes repeatedly and exhaustively, with no other considerations.

The second algorithm was deemed the Seed and Extend algorithm (see Figure 2-13). The Seed and Extend algorithm begins by aligning the highest scoring pair of nodes between the two input networks. Then from that point onwards, every new pair of nodes aligned must each interact with a pair of previously aligned nodes in the network, such that every pair of nodes aligned by Seed and Extend will add at least one aligned pair of edges to the alignment. Seed and Extend then continues adding newly aligned node pairs exhaustively.

Taken together, the Greedy and Seed and Extend algorithms serve as bounding algorithms for the behaviour of GreedyPlus, which we expect to fall somewhere in between the two, depending on the edge alignment weight. They serve to verify that both scoring components of the GreedyPlus algorithm, the node-based similarity scores and the edge-based edge alignment weight, are contributing to the performance of the algorithm.

Seed and Extend Algorithm

```
alignment ← null
bestPair ← null
for all u ∈ D1 ∪ B1, v ∈ D2 ∪ B2 do
    if Score(u, v) > Score(bestPair) then
        bestPair ← ⟨u, v⟩
    end if
end for

alignment ← alignment + bestPair
while ∃u, v, x, w s.t. u ∉ alignment and (u, x) ∈ E1 and v ∉ alignment and (v, w) ∈ E2
and ⟨x, w⟩ ∈ alignment do
    bestPair ← null
    for all u, v do
        if Score(u, v) > Score(bestPair) then
            bestPair ← (u, v)
        end if
        alignment ← alignment + bestPair
    end for
end while
```

Figure 2-13 – The Seed and Extend algorithm for IIN alignment. This is a simple algorithm for IIN alignment used as a comparison baseline for GreedyPlus. Beginning with a single node pair, it strictly follows the edges in both input networks as it aligns additional nodes and adds them to the existing sub-alignment.

2.5.3 Network Creation

The *S. cerevisiae* and *C. elegans* networks were created using interaction data from Tonikian et al. and Xin et al..[46,49] 24 SH3 domains were identified in *S. cerevisiae*, their 853 interactions were experimentally identified, and then 497 ligand targets for those interactions were predicted. Similarly, 33 SH3 domains, 466 SH3-mediated interactions, and 433 SH3 ligands were identified in *C. elegans*. Each SH3 domain was represented by an individual node, with the exception of the first two SH3 domains on the Sla1 protein, which were treated as a single domain in the original prediction procedure, as the two domains could not be purified separately. These networks cover approximately half of the SH3 domains in each species; the remainder failed for various experimental reasons and were thus excluded from our work.

Each predicted interaction included a target peptide sequence of length 15. When these peptide targets occupied non-overlapping positions in the target proteins, each peptide target was as an independent binding site represented as an independent node. When binding sites overlapped, they were merged, when possible, into a singular node representing no more than 30 amino

acids. Overlapping binding sites with a combined length of more than 30 amino acids were manually separated into multiple nodes with minimum sequence overlap between nodes.

The interaction networks for the other yeast species were similarly created using interaction data from Sun et al..[105] Each network contains approximately 500 predicted interactions, generated by using the 30 position weight matrices created for 24 *S. cerevesiae* SH3 domains and mapping them to each yeast proteome in which an orthologous SH3 domain exists. The networks for three species – *S. paradoxus*, *S. mikatae*, and *S. bayanus* – the only three datasets sourced from Fungal Comparative Genomics (original source: Kellis et al.[155]), were excluded due to unusual performance. In particular, pairwise alignments between the remaining 20 yeast species had an average protein orthology recovery rate of 56% and a minimum of 29%, compared to just 24% for alignments between the three excluded species and the remaining 20 yeast species.

2.5.4 Orthology Data

The orthology dataset for *S. cerevisiae* and *C. elegans* was created from a union of orthology mappings retrieved from Ensembl,[156] Inparanoid,[157] and OrthoMCL,[158] The orthology datasets for the yeast species were produced by Wapinski et al..[159]

2.5.5 Similarity Feature Data

We began with 29 similarity features for assessing every pair of nodes from two separate networks. Sequence similarity was calculated between every pair of proteins using BLAST-P, taking both the raw score and the coverage as features, as well as every pair of domains. Sequence similarity between ligand sites was calculated using the Smith-Waterman algorithm with the BLOSUM62 scoring matrix, as implemented by JAligner.[160]

Functional similarity was calculated between proteins using TCSS,[148] taking the biological processes, cellular components, and molecular function scores as three separate similarity features. Graphlet degree similarity was calculated between all SH3 domain and between all ligand site nodes separately, as described by Pržulj et al.,[124] as were the remaining 18 similarity features – betweenness, closeness, degree, eccentricity, neighbourhood connectivity, radiality, stress centrality, and topological coefficient. Raw values for these features were obtained using the NetworkAnalyzer plug-in in Cytoscape,[6,161] and a raw similarity value was calculated for

each pair of domains [i, j], $i \in D_1$, $j \in D_2$, $raw_{i,j} = \max(score_x - score_y) - (score_i - score_j)\ \forall\ x \in D_1, \forall\ y \in D_2$, and then normalized logarithmically to the interval [0,1] using the formula: $\forall\ x \in D_1, \forall\ y \in D_2$, $adj_{i,j} = \log(raw_{i,j}) / \log(\max(raw_{x,y}))$. Similarity scores between ligand nodes were calculated similarly.

2.5.6 Parameter Training Procedure

The parameter training procedure used was a random hill-climbing heuristic, designed to find parameter sets that maximized the orthologies found (RPOs). For each set of similarity features trained, we randomly generated a weight parameter in the interval [0,1], and generated a corresponding alignment. We then randomly incremented or decremented the first parameter by a step value if the new value would remain within the interval [0,1] and generated a new alignment. If the first alignment had more RPOs, then the parameter change was reversed and another parameter chosen to be incremented or decremented. Otherwise, the new parameterization was kept, and the parameter stepping repeated until no further improvement could be achieved. Then every other parameter would be retested for possible improvement via incrementation or decrementation.

This process was continued until no parameter could be either incremented or decremented to improve the orthology reproduction of the alignment produced. This procedure was iterated four times, using increasingly precise step sizes: $\sqrt[4]{0.01} \approx 0.32$, 0.1, 0.03, and 0.01, until convergence was achieved, resulting in a parameterisation at a presumed local maximum for orthology recovery.

For each set of similarity features used, the training procedure was iterated at least 5,000 times, producing at least 5,000 locally optimal parameter sets.

2.5.7 Similarity Feature Reduction

To reduce the full set of similarity features to a smaller set, redundant similarity features were identified by calculating Euclidean distances between each similarity feature matrix and performing principal component analysis. Similarity features that were highly similar to another feature (Euclidean distance with another similarity feature ≤ 0.10) were then removed, with preference given to removing the feature most similar to the other remaining features. The

following features were removed: BLAST coverage, TCSS cellular component, and TCSS biological process for proteins, betweenness, BLAST coverage, eccentricity, and radiality for domains, average shortest length path, betweenness, eccentricity, degree, radiality, and stress for binding site.

To further reduce the similarity feature set, GreedyPlus was re-optimized with the remaining 18 similarity features to identify features that could be removed without negatively impacting algorithmic performance. The similarity feature given the lowest weight in parameter sets associated with the top 50 results from the training procedure was identified as the most uninformative feature and removed. This process was repeated until an effective minimal set of similarity features was identified, whereby the removal of any additional feature resulted in loss of orthology recovery.

Chapter 3

Dynamics of Protein-Protein Interaction Conservation

I conceived of, designed, and conducted this project. Gary D. Bader supervised and advised this project.

Special thanks to Shobhit Jain for providing DoMoPred data, and Mark G.F. Sun for insight and commentary on network rewiring.

3 Dynamics of Protein-Protein Interaction Conservation

3.1 Introduction

Experimental efforts to map physical protein-protein interactions (PPIs) continue to produce increasing amounts of data, resulting in increasingly complete organismal protein-protein interaction networks (PPINs). These networks capture much information about complex biological systems, including protein function[122,162] and modular structure, representing biological pathways and protein complexes.[163]

An important way to understand biological systems is to study their evolution. Molecular evolution of biological sequences is well understood, but there is very little support for a general theory of network evolution. Current PPIN evolution models largely fall into categories: scale-free (preferential attachment), geometric, and duplication-divergence.[123,124] Generally, these models are evaluated based on their ability to simulate random *de novo* networks that bear similar graph theoretic statistics as real PPINs, rather than their explanatory power for network evolution.

There have been more direct efforts to measure and quantify PPIN evolution. Walhout et al. introduced the term *interolog* in 2000 as an analog to orthologs for PPIs; if two interacting proteins in the same species each have, in another species, orthologs that also interact, then those PPIs are interologs and presumed to be conserved.[114] There have been numerous efforts to estimate the rate of network rewiring, using various strategies including examining paralogs,[102] PWM scanning,[105] incorporating structural modelling,[109] and direct comparison, though using various counting mechanisms,[103,104,117] depending on the specific scientific perspective at work. Published conclusions about network rewiring rates have, however, varied widely, depending on factors like the species and the network regions assessed.

One application of PPIN evolution concepts is the development of network alignment algorithms. The principal assumption in network alignment research is that PPIs are conserved across species because PPIs define function and function is conserved. PPINs, then, can be aligned to each other, for a variety of scientific purposes, including function prediction, identifying homologs/orthologs, and investigating PPIN evolution. In particular, network alignment algorithms are divided into local, focusing on aligning small PPIN regions such as

pathways or complexes, and global, which attempt to align whole PPINs, approaches, with commensurately scaled objectives.

There are several other pervasive assumptions about PPIN evolution in the literature. For instance, it is typically assumed that PPI conservation between species is high, and thus PPIs can be transferred between species.[21,104,117,164,165] A related assumption is that highly conserved proteins are likely to have highly conserved PPIs, which then warrants the use of sequence similarity as a key information source for aligning nodes in network alignment algorithms.[70,100,166] However, our understanding of the relationship between sequence similarity and the appropriate alignment of proteins is limited. Whole protein BLAST scores are typically used to determine the suitability of a protein alignment, with little attention paid to other important evolutionary events. Gene duplication is particularly difficult to handle, as it changes the multiplicity of aligned proteins, and the divergence in function is expected to be matched by a divergence in interaction partners.[167] Similarly, gene fusion events should generate proteins with novel sets of interaction partners, possibly composites of the fused proteins' interaction partners. Within a single protein, there are other observable evolutionary phenomena that should generate predictable PPIN changes. Deletions, mutations or major rewriting of protein sequence sections involved in protein-protein interactions should subsequently lead to the loss of those interactions. Exon shuffling of those regions to other proteins may also result in the transfer of interactions from one protein to another.[168] Such events could easily have phenotypic effects of interest, though the relationship between PPI evolution and phenotype has not been well-established.[169]

Ideally, we would understand how protein-protein interactions are formed and controlled down to the residue level for most PPIs, but that remains difficult mainly due to the lack of detailed PPIN data across organisms. However, the modularity of protein domains may provide a useful starting point from which a deeper understanding of interactomes can be developed, and network evolution can be connected to protein and genome evolution.[47] Protein domains are conserved sequence, function and structure units found throughout the proteome and some mediate PPIs. The proliferation of domains has occurred in a variety of ways, including duplication of proteins containing domains, and domain shuffling.[170-173] Peptide recognition modules (PRMs) are a subset of domains that bind linear peptide motifs.[174] They are relatively well-studied, and since their primary molecular function is to mediate PPIs, they could be used to study the relationship

between protein evolution and interactome evolution. Just as function is conserved on the protein level, we would expect that function would also be conserved on the domain level;[175] we might speculate that conserved PRMs should also show conservation of interaction targets, and that duplicated PRMs should diverge in target interactors.

In this work, we investigate interaction conservation in PPINs, with an eye towards the consequences for network alignment research. We also consider several previously unconsidered evolutionary phenomena, such as changes in domain signature and gene duplication, and how they might be considered in network alignment. We find that current interactomic data show low rates of interaction conservation, especially for duplicated proteins, to an extent that has negative implications for global network alignment strategies. We further find that even within the limited scope of SH3-mediated PPIs, there is little evidence of PPI conservation. Combined with other research on gene multi-functionality and its impact on network analyses,[122] this work casts doubt on the utility of methods that presume that PPIs are highly conserved. The mechanisms of network evolution need to be better understood, and methods created that build upon these mechanisms, in order to devise better, more accurate PPIN alignment methods and better understand how biological systems evolve.

3.2 Results

3.2.1 Interolog conservation across species

Under either the preferential attachment or duplication-divergence models that dominate network evolution,[167,176,177] high-degree proteins should remain high-degree as they gain interactions at a faster rate than low-degree proteins. However, examining this in existing protein-protein interaction networks shows that this effect is not strong. Comparing orthologs and their degrees in the PPINs of five model species (human: *H. sapiens*, mouse: *M. musculus*, fruitfly: *D. melanogaster*, worm: *C. elegans*, and budding yeast: *S. cerevisiae*), the correlation between the degrees of orthologs is very low (see Figure 3-1 and Table 3-1). While there are known biases and incompleteness in PPI data, this result suggests that for PPIN alignment, the presence (or absence) of aligned interactions is an unreliable signal from which to infer orthology.

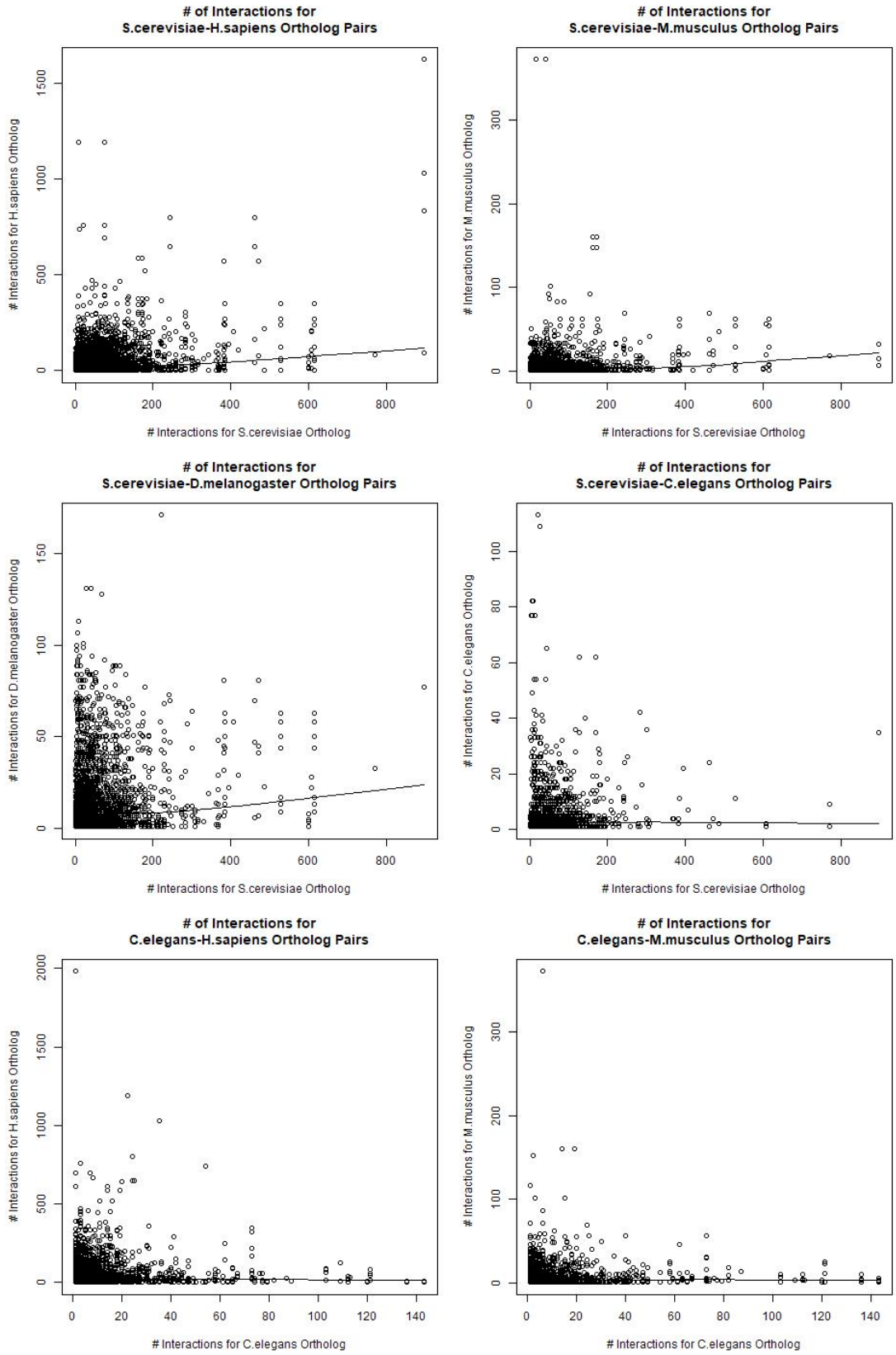


Figure 3-1 – Comparisons of the number of interaction partners for all pairs of protein orthologs in various pairs of species. (Continued on next page.)

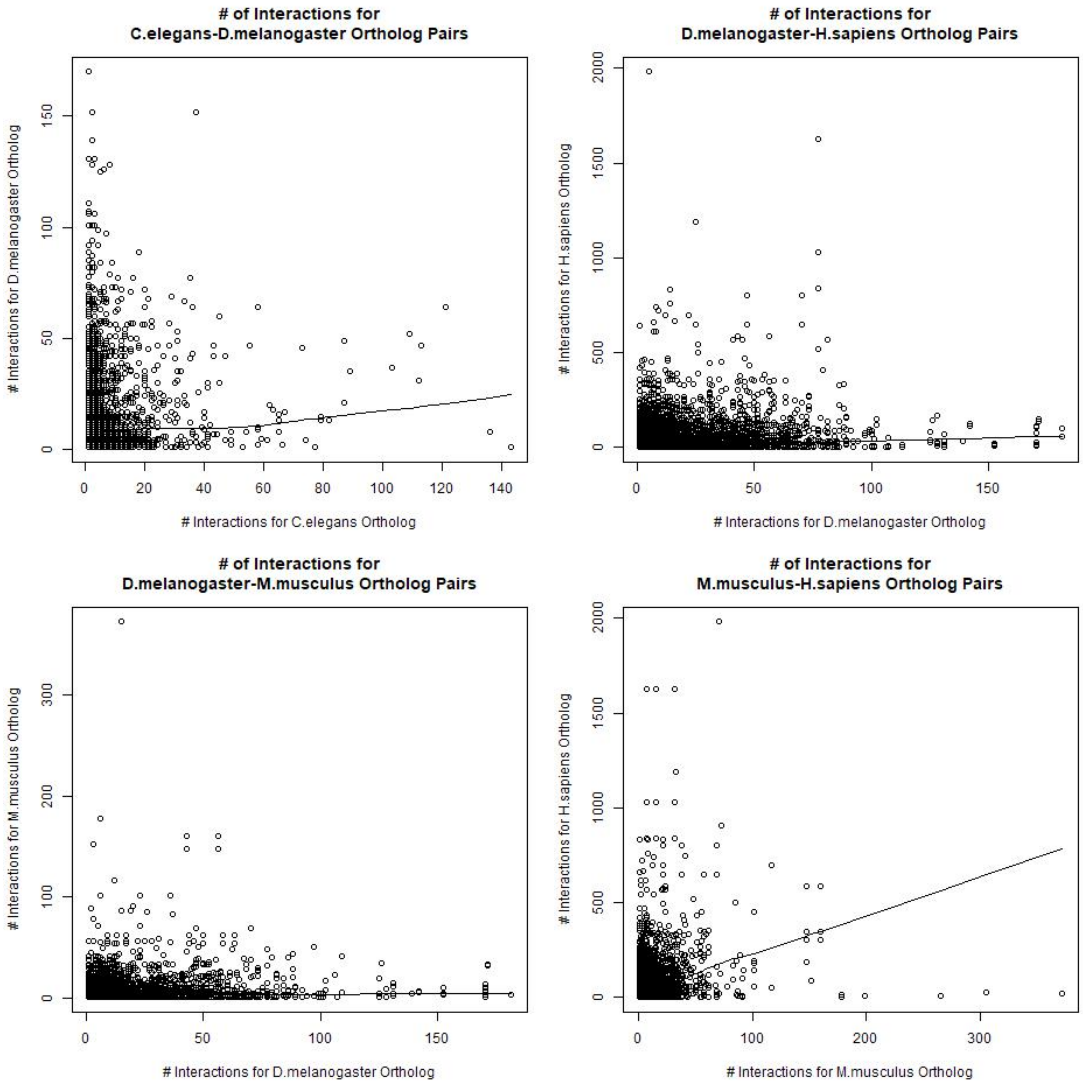


Figure 3-1 – Comparisons of the number of interaction partners for all pairs of protein orthologs in various pairs of species. The x- and y- axes have different scales to account for different proteome and interactome sizes between species. A LOWESS (locally weighted scatterplot smoothing) curve is included as an approximate line of best fit.

Correlation	*Mouse*	*Fruitfly*	*Worm*	*Yeast*
Human	0.2775	0.1554	0.0668	0.3477
Mouse		0.1978	0.0541	0.1727
Fruitfly			0.1269	0.2165
Worm				0.0748

Table 3-1 – Correlation coefficients between the number of interaction partners for all pairs of protein orthologs in various pairs of species.

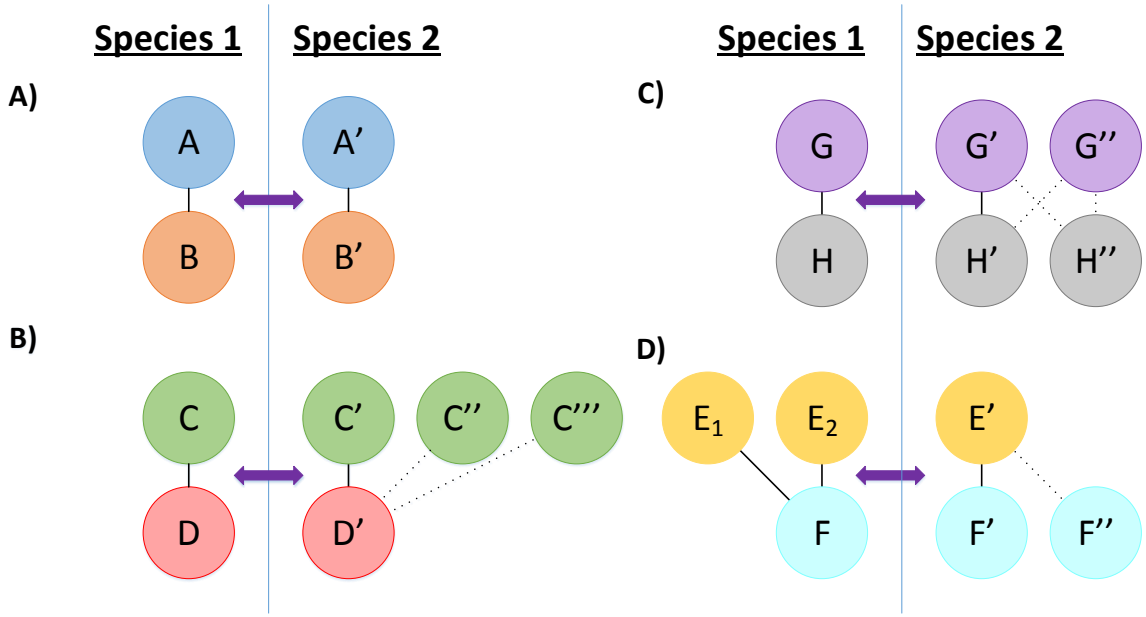


Figure 3-2 – Illustration of various interolog scenarios. Interologs occur when an interaction exists between two proteins, A and B, and an interaction exists between the orthologs of those two proteins in another species, A' and B'. These two interactions are interologs of each other, and are presumed to be conserved. However, counting interologs between two interactomes can be complicated by diverse genomic evolutionary events, such as gene duplications. In B), the protein C in species 1 has been duplicated twice, resulting in orthologs C', C'', and C''' in species 2, but only C' and D' interact, with dotted lines representing the missing interactions from D' to C'' and C'''. In this case, it is ambiguous if the interaction between C and D is conserved: out of three possibilities for an interolog to appear in species 2, only one such interolog appears. In C), both proteins G and H have been duplicated in species 2, but only G' and H' interact. Thus, the interaction exists between one pair of orthologs, but not between the three other possible pairings of G and H orthologs. In D), gene duplication has occurred in both species, with two interactions between both E1 and E2 with F in species 1 only conserved once between E' and F'. Thus, the level of interolog conservation varies depending on which direction of comparison is of interest, from species 1 to species 2 or vice-versa.

We next examined the rate of conservation in the PPINs of different species by examining interolog frequency. Some complexities of interolog definition must first be addressed. As protein orthologies are not always one to one, due to protein duplication and deletion events, interologs can be counted in different ways, presenting different perspectives on how frequently interactions are conserved (see Figure 3-2). Previous work on interolog mapping has used reciprocal best BLASTP hits to identify orthologs instead of established orthology databases, as a means to limit false positives.[18,104,117] While this approach may be sensible for the purpose of predicting PPIs in new species, it is not suitable for use in PPIN evolution studies because it misses a large set of orthologs that have lower BLASTP similarity scores. As we are interested in estimating rates of interaction conservation, we need to consider all potential interologs, rather

than just the subset considered likely to have been conserved (see Figure 3-3). To support this, we define orthologs more liberally by collecting them from multiple ortholog definition services.

Based on available data, interologs are rather uncommon. In Table 3-2, we show the rates at which any potential interolog can be found in a target species for a given pair of interacting proteins in another origin species. We examined every known protein-protein interaction (PPI) in the origin species, taking both interactors, identified their respective orthologs in the target species, and then checked for any interaction between any two members from the two distinct ortholog sets. We ignore *unmatchable* interactions, interactions for which one or more of the original interactors had no orthologs in the target species, as these interactions cannot be conserved due to larger genomic changes. Instead, we consider only *matchable* interactions, interactions for which there are orthologous partners to their interactors in the target species, and whether they were matched at all by at least one interaction between any of the interactors' corresponding orthologs in the target species.

These values represent how often a PPI in the origin species is conserved in the target species. The values fluctuate greatly, due to differing interactome sizes and ortholog complements, as well as incompleteness in the interactome data, but even when the largest interactome, *H. sapiens* (human), is the target, fewer than 37% of the interactions from any other species were found to have interologs. Thus, the majority of matchable interactions are not conserved given current data. Furthermore, while many sequence alignment methods rely on protein sequence similarity to achieve high biological quality (see 1.2.4.1 Biological Assessment and 2.3 Discussion), as shown in Table 3-3, sequence similarity is not a good indicator for the conservation of an interaction. By keying too heavily on biologically meaningful node alignments, network alignment methods may be sacrificing biologically meaningful edge alignments, tacitly subordinating the alignment of interactomes to the alignment of the proteomes.

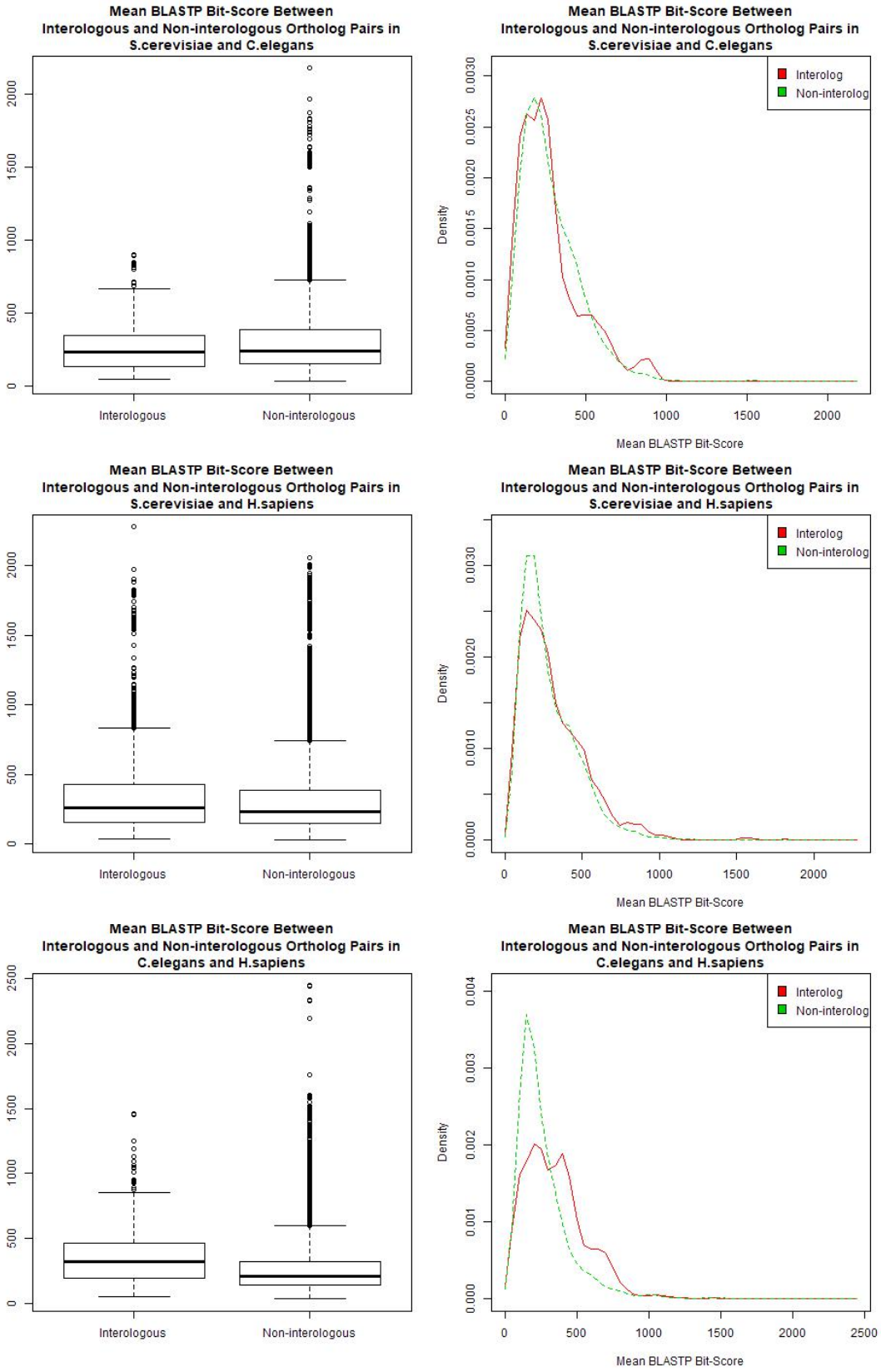


Figure 3-3 – Boxplots and density plots comparing the mean BLASTP bit-score between interologous and non-interologous protein pairs between selected pairs of species.
(Continued on next page.)

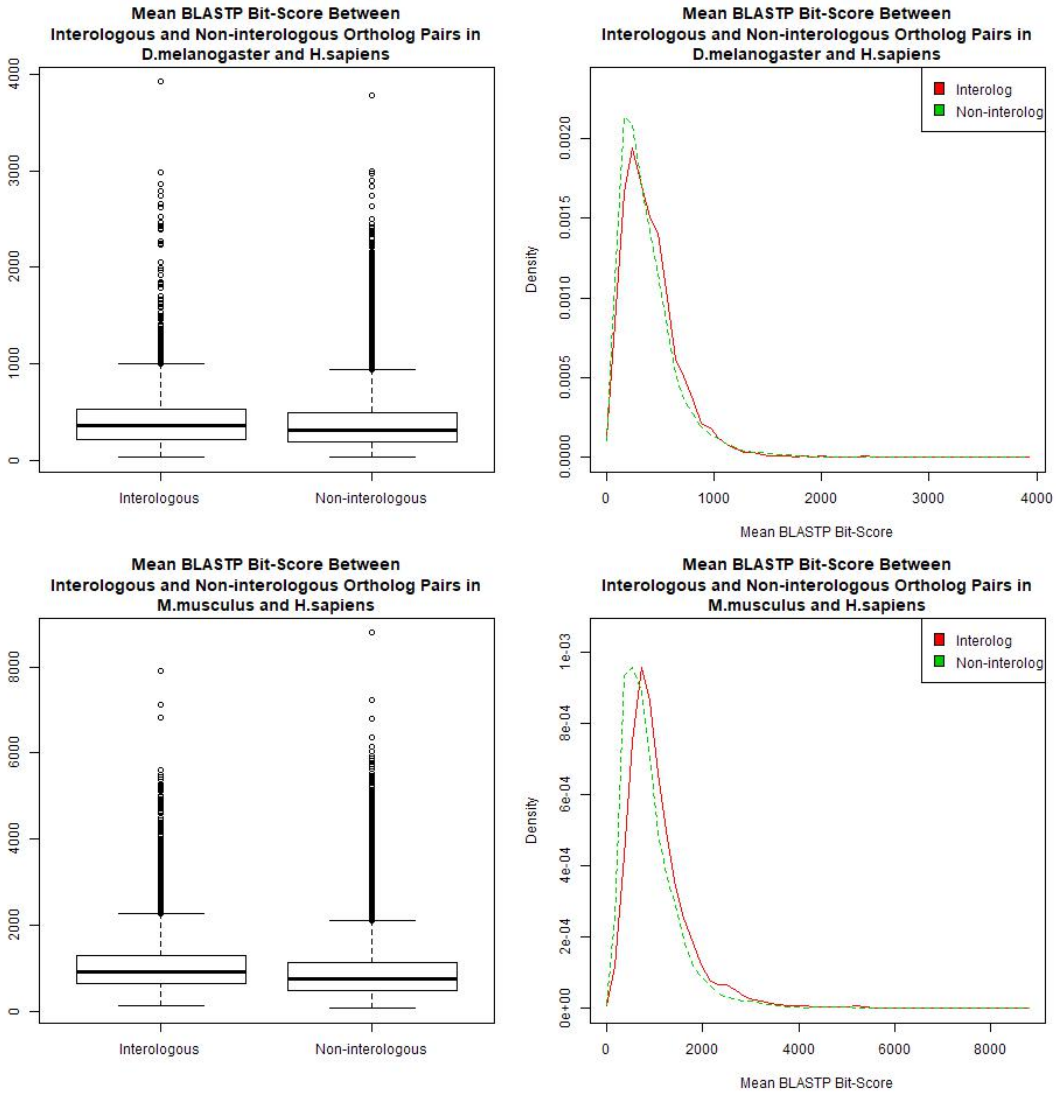


Figure 3-3 – Boxplots and density plots comparing the mean BLASTP bit-score between interologous and non-interologous protein pairs between selected pairs of species. For each pair of interacting proteins in the first species, we examined every corresponding pair of orthologs in the other species, identifying them as interologs if the orthologs interacted or non-interologs otherwise. We then used BLASTP on every pair of orthologs, then for each interolog/potential interolog calculated the mean bit-score between the two constituent ortholog pairs, then plotted these scores. While there are differences in the two distributions of BLASTP bit-scores visible in the *C. elegans – H. sapiens* and *M. musculus – H. sapiens* density plots, there is substantial overlap between the BLASTP bit-scores of the interologous and non-interologous ortholog pairs in all the above plots, particularly at the lower end of bit-scores.

		Target Interolog Species				
		Human	*Mouse*	*Fruitfly*	*Worm*	*Yeast*
Origin Interaction Species	*Human*		6578/99477 (6.61%)	5036/94277 (5.34%)	1219/38187 (3.19%)	8326/43021 (19.35%)
	Mouse	5392/14921 (36.14%)		1205/8757 (13.76%)	312/4395 (7.10%)	912/2665 (34.22%)
	Fruitfly	3182/22222 (14.32%)	852/12784 (6.66%)		354/6644 (5.33%)	1948/8867 (21.97%)
	Worm	702/4660 (15.06%)	212/2904 (7.30%)	329/3447 (9.54%)		277/1119 (24.75%)
	Yeast	5702/40201 (14.18%)	847/26875 (3.15%)	1841/35425 (5.20%)	314/14358 (2.19%)	

Table 3-2 – Interaction conservation between five model organism species (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *S. cerevisiae*). For every protein-protein interaction in the origin species (left), we searched for corresponding interologs in each of the other four target species (top). Presented are the number of interactions in the origin species with at least one interolog in the target species, the number of interactions in the origin species that could have interologs in the target species based on available orthologs, and the ratio between them. Due to gene deletion, "de novo" gene creation, and undiscovered orthology relationships, not all proteins have orthologs in all other species. Thus some interactions are *unmatchable*, if one or more of their interactors do not have orthologs in a given target species. Only the *matchable* interactions, those whose interactors each have at least one ortholog in the other species, were counted in the denominators above. Of those interactions, those for which at least one interolog was found in the target species were counted in the numerator. Consequently, both proteome size and interactome size/coverage are key determinants of the background sizes shown in the above table.

Most interolog analysis focuses on conserved PPIs, but it is also important to consider how often an interolog is not found where it might be expected, as this helps us understand interactome evolution and how we should model network changes. In particular, when interactors are duplicated, the interaction between them may be conserved between each of the resulting duplicates, between some of them, or between none of them. To study this, we repeated the above analysis, but asked how often an interaction was conserved out of all opportunities for it to be conserved among all the orthologs of the interactors.

We found that when there are multiple opportunities for an interaction to be conserved, it is decreasingly likely that each possible interolog is found. While ~36% of mouse PPIs have at least one interolog in human, mouse PPIs are, on average, only conserved ~26% of the time amongst all the possible ortholog pairs in human (see Figure 3-4, see Figure A-1 for all species comparisons). Furthermore, we found that of all the human PPIs that could be conserved from the mouse interactome, only 15% were found in the human interactome. This difference exists because highly duplicated proteins create many more opportunities for potential interologs, thus

numerically dominating the calculation, and the more often a protein has been duplicated, the less likely its interactions are to be conserved (see Figure 3-5, see Figure A-2 for all species comparisons). This is consistent with duplication-divergence models of network evolution[129,167] and the hypothesis that PPIs correlate with a protein's function: when genes are duplicated, they will diverge in function, which manifests as rewiring of the interactions involving the duplicated proteins.[128]

As most network aligners are one-to-one, aligning each node to at most one other node, it is unclear how they treat gene duplications. However, many-to-many network aligners specifically will have to consider that the higher the multiplicity of a given node alignment, the less likely they will be aligning edges, and conversely, the more edges they align together, the less likely they'll be aligning paralogous proteins. This will create a conflict between two presumed goals for these aligners: to align together related gene families and to align PPIs at a high rate. Similarly, interolog mapping efforts may need to pay special mind to duplicated proteins. Comparing the average and overall rates of comparison listed in Figure 3-4 and Figure A-1 shows the outsized effect of gene duplication: even if the average PPI in the target species might be conserved at a reasonable rate, the relatively few PPIs with highly duplicated interactors can dramatically reduce the overall precision of interolog mapping, by a factor of up to half.

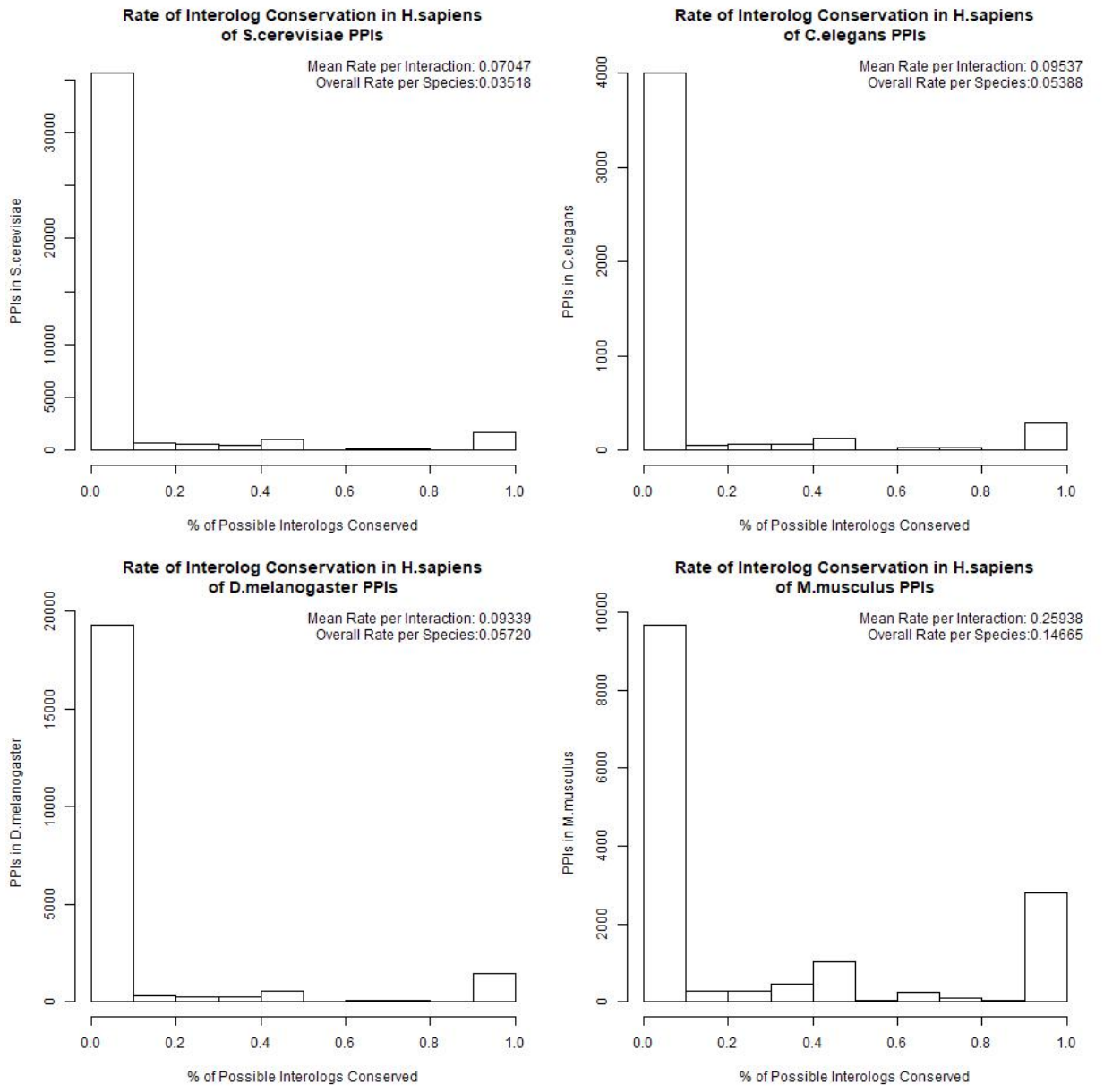


Figure 3-4 – The rate of conservation for PPIs from different species in the *H. sapiens* interactome. This figure shows the distribution of conservation rates per PPI in the origin species, where a value of 1.0 indicates that the PPI is conserved between *all* human orthologs of the protein interactors and 0.0 indicates that no interaction is ever found between any of the human orthologs. Compared to Table 3-2, which presented the % of PPI that had *any* conserved interolog in the human interactome, these rates of conservation are much lower, as there are often multiple human orthologs for any given protein in the origin species. In the top right are listed the mean rate of conservation for a given PPI in the origin species amongst all possible interologs in the *H. sapiens* interactome, and the total rate of conservation of all possible interologs in *H. sapiens*.

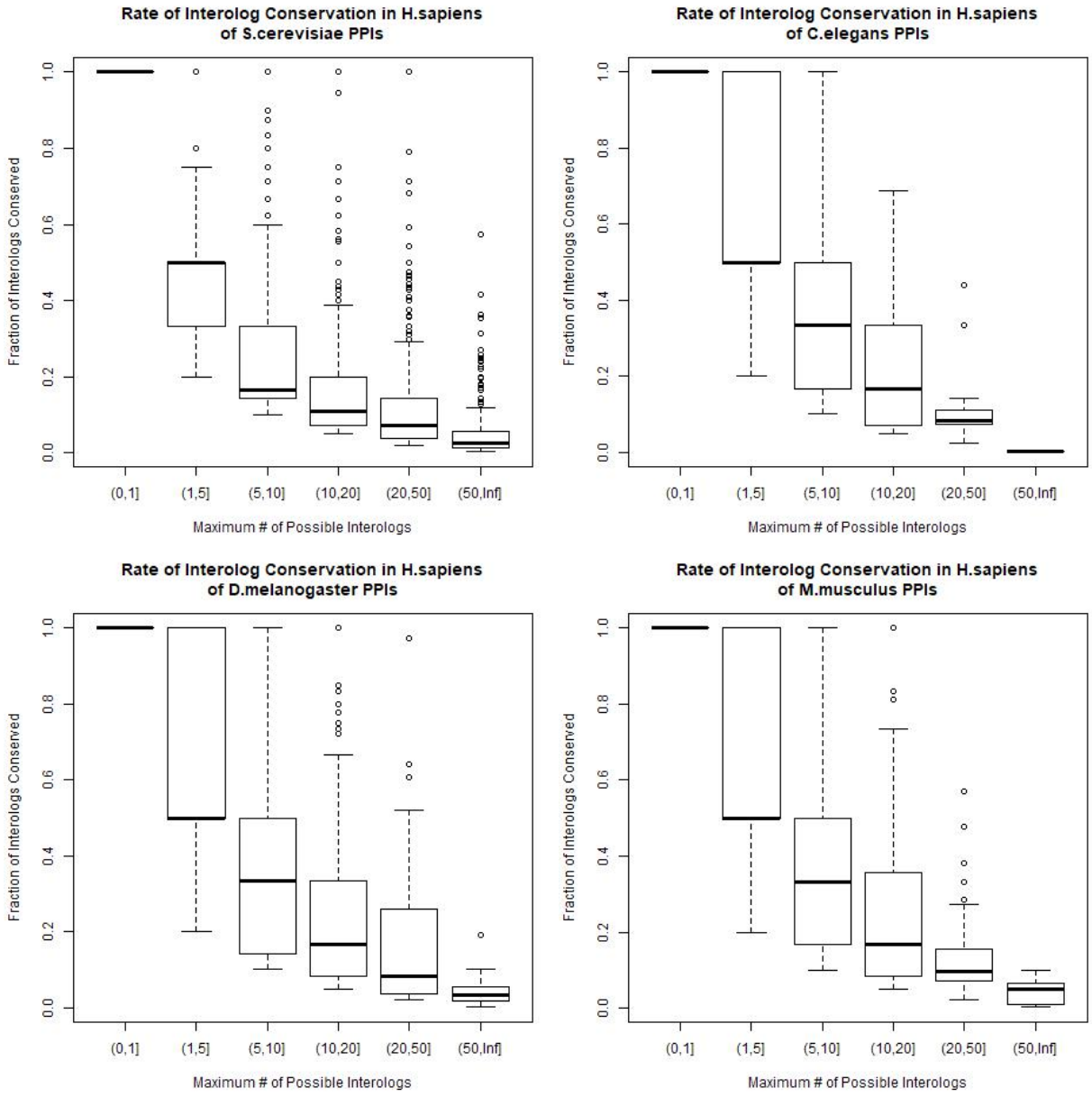


Figure 3-5 – % interolog conservation for PPIs in the *H. sapiens* interactome, grouped by the number of potential interologs. Due to gene duplication events resulting in many-to-many ortholog relationships, some PPIs can be conserved more than once, resulting in multiple potential interologs. These histograms show the level of interolog conservation of PPIs from various species in the human interactome, grouped by the maximum number of potential interologs that could have been found based on the number of orthologs in the human genome. PPIs with no interologs were excluded, as they overwhelmed all other PPIs due to low overall rates of interolog conservation.

3.2.2 Impact of PRMs on Protein Interactivity

The dynamics of network evolution are complex, with distinct contributions from diverse molecular-level evolutionary mechanisms that affect gene and protein sequences. Given that phenotypic variation is fundamentally driven by mutations at the gene level, gene and protein sequences could serve as indicators for evolutionary events at the network level, and there has been some work that has demonstrated this.[178] The larger problem of how to infer network evolution events from genetic sequence events remains unsolved, however.[47]

One way to explore this question is through peptide recognition modules (PRMs), which are a natural bridging point between evolutionary events at the sequence level and at the network level. As PRMs often exhibit highly specific but diverse binding profiles, even within the same PRM family,[50] it may be expected that proteins with more PRMs would be involved in more PPIs. We explored this correlation for a set of well-studied PRMs across species (see Table 3-3). While the number of interactions involving a given worm or yeast protein correlates well with the number of PRM domains, these figures fall off significantly for human, mouse, and fruitfly proteins. A similar pattern is seen with protein binding domains in general. This pattern is strongest with SH3 domains, the oldest of the analyzed PRM families, while the younger PRM families, such as SH2, PDZ, and WW, retain relatively high correlations in the two mammalian species. Also, Bromo domains generally have a strong negative correlation.

To further examine these effects, we compared the differences in the number of PPIs and the number of SH3 domains between orthologous protein pairs in which at least one of the orthologs has an SH3 domain (see Figure 3-6). We focused on SH3 domains for this analysis as they are the most numerous and best studied, however we also performed this analysis generally for all PRM domains collectively. (See 3.5.2 Domain Data for details.) Again, we generally observe a weak positive correlation on the number of PPIs involving a protein as the number of SH3 or PRM domains within that protein changes, especially for yeast orthologs. However, this pattern does not hold for the other species.

The differences in correlation patterns we observe between species and between domain types suggest multiple distinct evolutionary processes affecting the interactome. For example, domain expansion in larger mammalian proteomes may cause a stronger negative selection effect on shared interactors, like short proline-rich sequences targeted by SH3 domains.[179,180] Or domain

duplication/divergence may cause domains to expand in number within a protein and diverge in function to add functionality to a protein through evolution of new binding partners.[46] Another possibility is that specific biological functions evolve in different ways depending on selection pressure. SH3 domains, for example, are functionally enriched in endocytosis and other functions related to movement of the cellular boundary.[46] The difference in correlation between interactions and SH3 domains may reflect a change in how those functions, and the parts of the interactome responsible for those functions, have evolved before and within Arthropoda. This difference also suggests there's an interplay between PPIs and domain evolution,[175,181,182] and that interactomic concepts such as network alignment and interface-interaction networks[44,183] could provide additional insight into the evolution of proteins and species. With other work indicating that different regions of PPINs exhibit distinct local network topology,[113] PPINs may be better considered as collections of various subnetworks, each with its own functional or evolutionary processes in use, rather than as a uniform whole.

Correlation	All	Protein-binding	SH3	Bromo	PDZ	SH2	WW	PRM
Human	0.044	0.028	0.102	-0.141	0.091	0.133	0.256	0.144
Mouse	0.013	0.011	-0.016	-0.134	0.143	0.246	0.251	0.084
Fruitfly	0.013	-0.021	0.020	-0.042	-0.009	-0.174	-0.250	-0.004
Worm	0.048	0.166	0.430	*-0.660*	0.155	-0.126	*0.485*	0.325
Yeast	0.085	0.172	0.393	-0.525	*-0.803*	*N/A*	*0.718*	0.214

Table 3-3 – Correlation coefficients between the number of domains in a protein and the number of PPIs involving that protein, aggregated per species, using different domain types. Proteins with zero of a given domain type were excluded from the respective calculation. N/A indicates that no correlation was computable, because every protein in the specified species with the specified domain type all contain the same number, or none, of that domain. Italicized text indicates that the correlation coefficient was computed using fewer than 10 separate data points (proteins). The PRM grouping is an aggregation of all the specified PRM types: Bromo, FHA, GYF, PDZ, Polo-box, PTB, SH2, SH3, and WW (see 3.5 Methods).

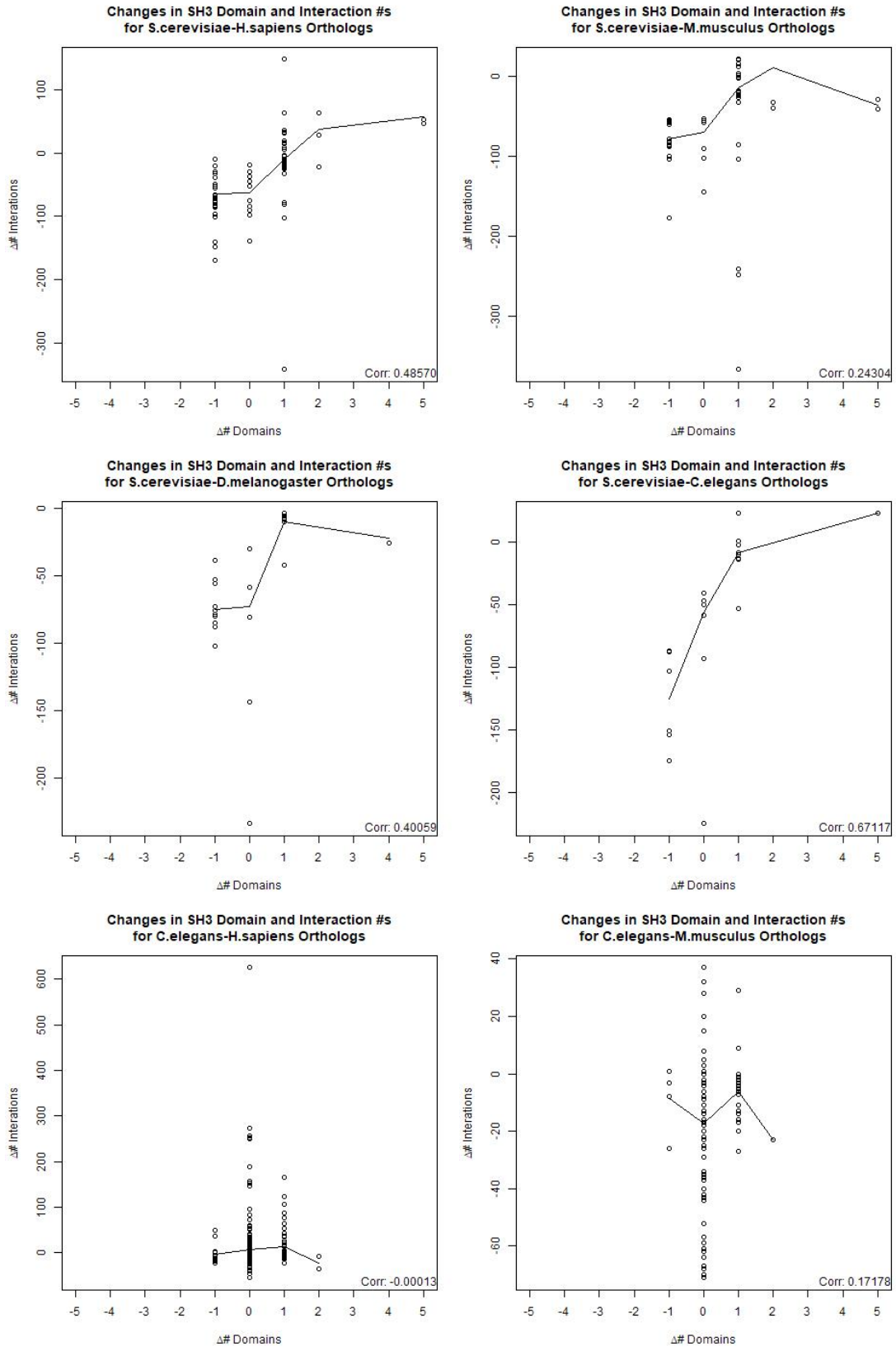


Figure 3-6 – Scatterplots relating the number of SH3 domains and the number of PPIs for orthologous protein pairs across different species. (Continued on next page.)

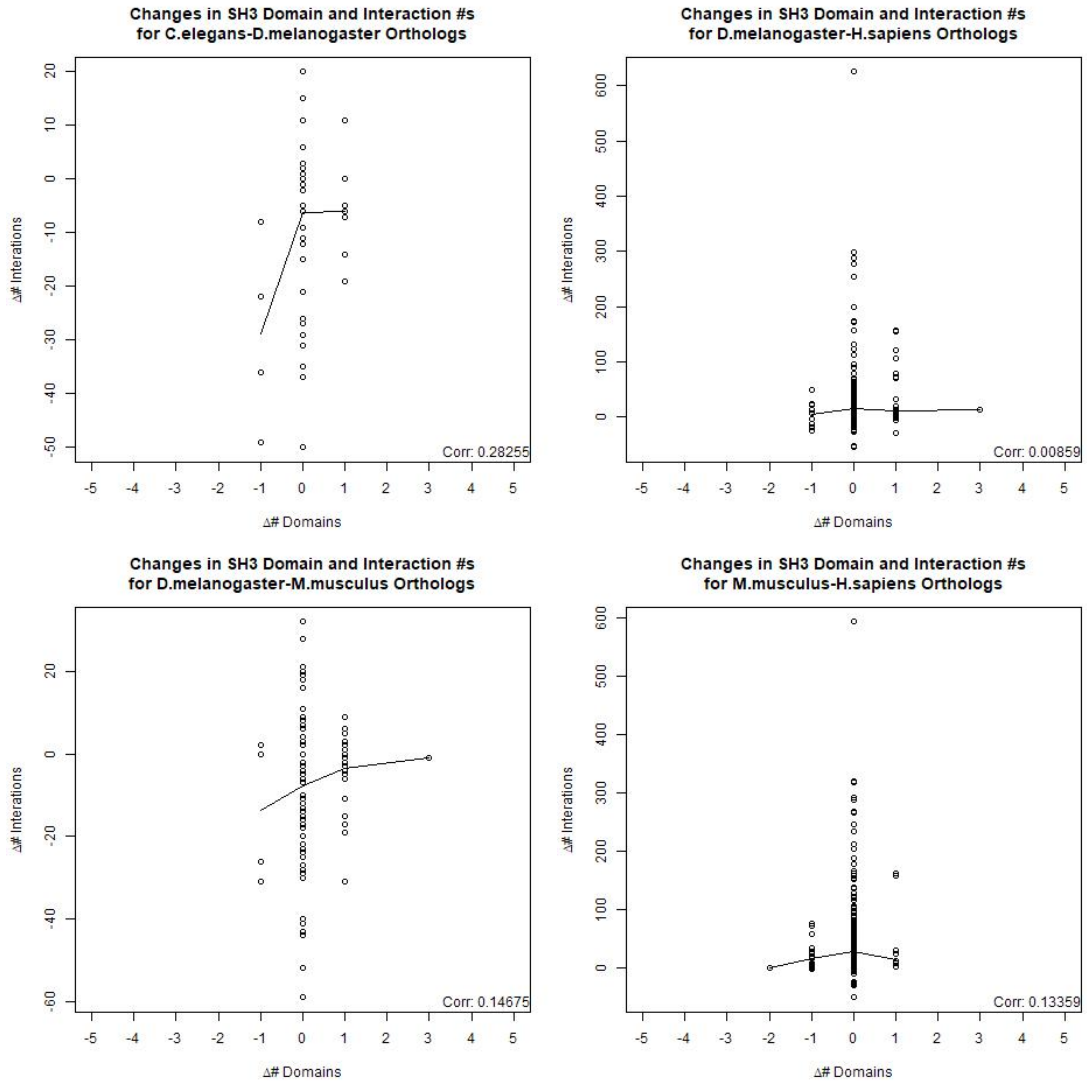


Figure 3-6 – Scatterplots relating the number of SH3 domains and the number of PPIs for orthologous protein pairs across different species. In each plot, each data point represents a pair of orthologous proteins between two species. Plotted along the x-axis is the change in the number of SH3 domains from the ortholog in the first species, as per the plot title, to the ortholog in the second species. Plotted similarly along the y-axis is the change in the number of PPIs for each ortholog. A LOWESS curve indicates an approximate line of best fit and a correlation coefficient displayed in the bottom right. Orthologous protein pairs with the same number of the specified domain are displayed visually, but were excluded from the correlation coefficient calculation, because of their abundance. Note that while the x-axis is standardized, the y-axis scale varies, reflecting differences in interactome density. Note also that the data points at Δ#SH3 Domains ≥ 2 both represent orthologies involving intersectin-1 or intersectin-2, which are highly similar paralogs with five SH3 domains in *H. sapiens*.

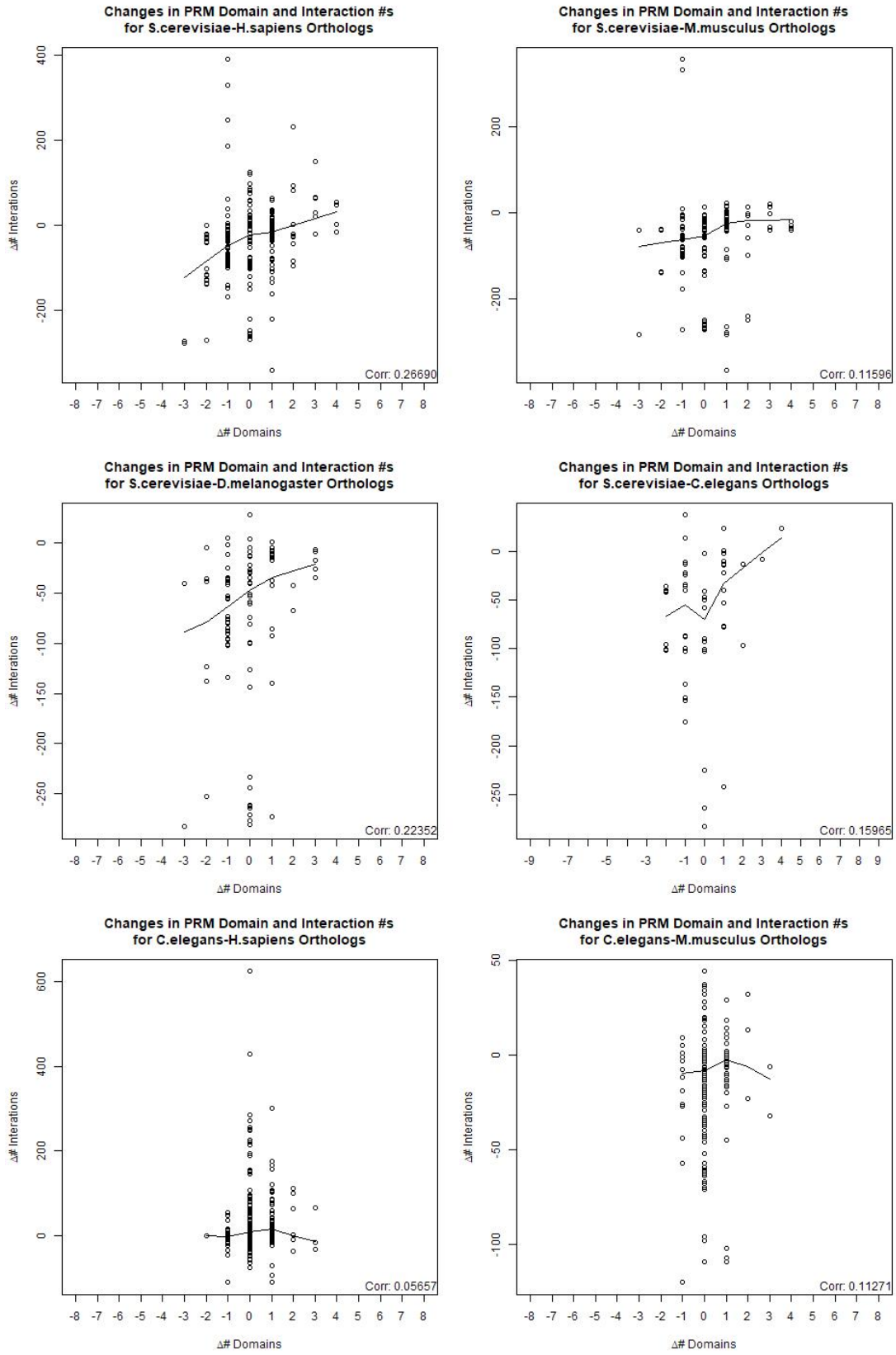


Figure 3-7 – Scatterplots relating the number of PRMs and the number of PPIs for orthologous protein pairs across different species. (Continued on next page.)

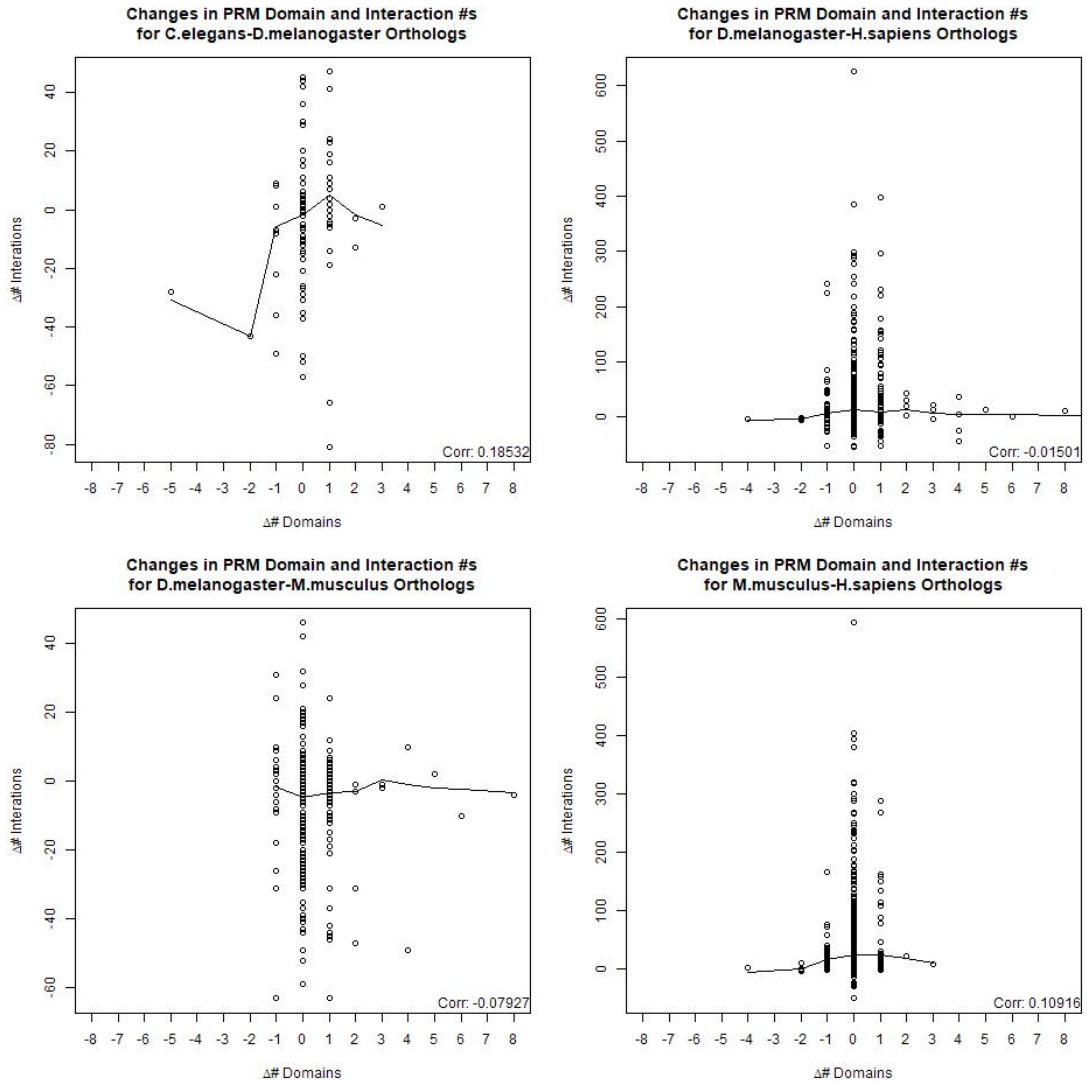


Figure 3-7 – Scatterplots relating the number of PRMs and the number of PPIs for orthologous protein pairs across different species. In each plot, each data point represents a pair of orthologous proteins between two species. Plotted along the x-axis is the change in the number of PRMs from the ortholog in the first species, as per the plot title, to the ortholog in the second species. Plotted similarly along the y-axis is the change in the number of PPIs for each ortholog. A LOWESS curve indicates an approximate line of best fit and a correlation coefficient displayed in the bottom right. Orthologous protein pairs with the same number of the specified domain are displayed visually, but were excluded the correlation coefficient calculation, because of their abundance. Note that while the x-axis is standardized, the y-axis scale varies, reflecting differences in interactome density.

3.2.3 The importance of domain architecture in protein-protein interaction conservation

To examine specific examples of sequence changes that are likely to have an important effect on the PPIN, we examined whether any of the 214 human SH3 proteins have a different number of SH3 domains in yeast, worm, fruitfly, or mouse orthologs, indicating that a domain gain or loss event has occurred at some point along the lineage. Beginning with a computationally generated list of these events, we manually verified each individually, examining the larger homologous protein family, verifying homology and domain identification by examining homologs from other species such as *R. norvegicus* and *D. rerio*, and expanding/extending the lineages as appropriate. While most SH3 protein orthologs have the same number of SH3 domains within ortholog groups, many SH3 domain number changes within the gene families were found (see Table 3-4). When one ortholog has an SH3 domain absent in the other, we expect that the second ortholog would participate in fewer SH3 domain-mediated interactions, resulting in a PPIN change that identifiable based on a change at the sequence level.[47] Thus, even within ortholog groups, expected to conserve function between species, we observe important sequence changes likely to affect the PPIN.

To further probe the effect of sequence changes on PPINs, we examined the intersectin gene family, which is the best studied multi-domain family with domain number changes between species. Human intersectin-1 (ITSN1) contains five SH3 domains, the most of any human protein. It has a less well studied paralog ITSN2 and orthologs in yeast, worm, and fruitfly, though the fruitfly ortholog, DAP160, has only four SH3 domains while yeast ortholog EDE1 has none. Human ITSN1 is known to have two isoforms, a short-form of 1220 amino acids and a brain-specific long-form with 1721 amino acids, both of which contain all five SH3 domains.[184,185] Both human ITSN1 and human ITSN2 are implicated in endocytosis and exocytosis, possibly as a regulator of the formation of clathrin pits.[186,187]

Table 3-4 – Evolutionary gain/loss events of SH3 domains.

	Yeast Protein		Worm Protein		Fruitfly Protein		Mouse Protein		Human Protein	
	Name	# SH3 Domains	Name	# SH3 Domains	Name	# SH3 Domains	Name	# SH3 Domains	Name	# SH3 Domains
Domain Gain Events	DGR2	0	SYM-4	0			AHI1	1	AHI1	1
	YMR102C	0								
	RGA1	0	TAG-325	0			ARHGAP12	1	ARHGAP12	1
	RGA2	0								
			Y44E3A.4	2	CINDR	3	CD2AP	3	CD2AP	3
							SH3D21	3	SH3D21	3
							SH3KBP1	3	SH3KBP1	3
							GAS7	0	GAS7	1
			CGEF-1	0	CG30440	0	MCF2	0	MCF2	0
							MCF2L	1	MCF2L	1
									MCF2L2	0
			AAP-1	0	PI3K21B	0	PIK3R1	1	PIK3R1	1
							PIK3R2	0	PIK3R2	0
							PIK3R3	1	PIK3R3	1
	HMT1	0			ART8	0	PRMT2	1	PRMT2	1
	BUD2	0	GAP-3	1	VAP	1	RASA1	1	RASA1	1
			GAP-1	0			RASA2	0	RASA2	0
							RASA3	0	RASA3	0
							RASA4	0	RASA4	0
									RASA4B	0
							RASAL1	0	RASAL1	0
	MDR1	0	TBC-18	1	CG12241	1	SGSM3	1	SGSM3	1
			SHN-1	0	PROSAP	1	SHANK1	1	SHANK1	1
							SHANK2	1	SHANK2	1

							SHANK3	1	SHANK3	1
	PFA3	0	DHH-6	0	CG4483	1	ZDHHC6	1	ZDHHC6	1
					CG5196	0				
Domain Loss Events			CSK-1	1	CSK	0	CSK	1	CSK	1
							MATK	1	MATK	1
			MOM-4	0	DSTYK	0	DSTYK	0	DSTYK	0
			W03A5.1	1						
			Y105C5A.24	0						
	BZZ1	2	TOCA-1*	1	CIP4	1	FNBP1	1	FNBP1	1
			TOCA-2*	1			FNBP1L	1	FNBP1L	1
							TRIP10	1	TRIP10	1
			SEM-5	2	DRK	2	GRAP	2	GRAP	2
									GRAPL	1
							GRAP2	2	GRAP2	2
							GRB2	2	GRB2	2
			NCK-1	3	DOCK	3	NCK1	3	NCK1	3
							SH3TC1	1	SH3TC1	1
							SH3TC2	2	SH3TC2	1
	CDC24	0	VAV-1	2	VAV	1	VAV1	2	VAV1	2
							VAV2	2	VAV2	2
							VAV3	2	VAV3	2
							NCK2	3	NCK2	3
Multiple Domain Gain/Loss Events			UNC-73	1	TRIO	1	ARHGEF25	0	ARHGEF25	0
							KALRN	2	KALRN	2
							TRIO	2	TRIO	2
	EDE1	0	ITSN	5	DAP160	4	ITSN1	5	ITSN1	5

							ITSN2	5	ITSN2	5
					CG43729	1	STAC	2	STAC	1
							STAC2	2	STAC2	1
							STAC3	2	STAC3	2
			ARK-1	1	ACK-LIKE	0	TNK1	1	TNK1	1
			SID-3	0	SHARK	0	TNK2	1	TNK2	1

Table 3-4 – Evolutionary gain/loss events of SH3 domains. Listed are all the human proteins with SH3 domains, for which there exists a mouse, fruitfly, worm, or yeast ortholog or a paralog with a different number of SH3 domains. They are organized into apparent phylogeny, based on MUSCLE[4] and Ensembl[156] data, then further grouped by whether there is an apparent SH3 gain or loss event, or some combination of multiple gain/loss events. *The evolutionary relationship between worm TOCA-1 and TOCA-2 and with worm/human FNBP1, FNBP1L & TRIO is unclear, possibly reflecting distinct gene duplication events or a combination of duplication and deletion events.

For each of the intersectin orthologs in these species, we gathered their PPIs from multiple sources and counted the number of interologs between each pair of proteins (see Table 3-5). The PPI data is inconsistent between species, with the fruitfly and mouse orthologs having unusually few protein-protein interactions, which possibly reflects data incompleteness rather than true interaction sparseness.

	# of PPIs	# Interologs with				
		Yeast EDE1	Worm ITSN-1	Fruitfly DAP160	Mouse ITSN1	Mouse ITSN2
Yeast EDE1	117					
Worm ITSN-1	66	5 *(25)*				
Fruitfly DAP160	16	2 *(9)*	1 *(8)*			
Mouse ITSN1	13	2 *(6)*	5 *(11)*	2 *(5)*		
Mouse ITSN2	1	0 *(0)*	0 *(1)*	0 *(1)*		
Human ITSN1	96	13 *(46)*	17 *(53)*	10 *(15)*	11 *(15)*	0 *(1)*
Human ITSN2	88	8 *(38)*	8 *(51)*	6 *(13)*	5 *(14)*	0 *(1)*

Table 3-5 – Number of interologs shared between intersectin orthologs. The number of potential interologs for each comparison varies depending on the conservation of protein orthologs in each species. Thus, for each protein in each pair of orthologs, we counted the number of protein interactors that have orthologs in the other species, then take the lower of the two and report it in parentheses. This value should be considered an approximation of how many interologs are theoretically possible, as protein duplications cause a single interaction in one species to serve as an interolog for multiple interactions in the other.

To further break down these interologs and understand the nature of PPI conservation, we incorporate data on the binding targets of SH3 domains in yeast, worm and human, which specifies what proteins, and which regions on these proteins, are targeted by various SH3 domains. The yeast and worm SH3 binding data were generated with phage display, then verified with yeast-two-hybrid screens,[46,49] whereas the human SH3 binding data is predicted data created using DoMoPred[188] with PWMs created from phage display experiments,[50] using a 95% confidence cut-off. By attributing PPIs to the specific domains and ligands that are interacting, we hope to identify how gain/loss of domains and changes in domain/ligand specificity affect interaction conservation at the protein level. The PPIs for each protein are separated based on the

SH3 domain that mediates the interaction (see Table B-1), with PPIs not mediated by an SH3 domain listed separately.

Of the five yeast EDE1 interactions conserved in worm ITSN-1, none of them are mediated by the four worm SH3 domains for which we have binding data (see Figure 3-8). However, two EDE1 interactions, not including self-interactions, have human interologs mediated by human ITSN1's SH3 domains – EDE1 - HRR25 is conserved as ITSN1 - TTBK2 and EDE1 - PKC1 is conserved as ITSN1 - PKN3, both mediated by ITSN1's fifth SH3 domain. Four EDE1 interactions have human interologs mediated by human ITSN2's SH3 domains: EDE1 - AKL1 as ITSN2 - AAK1, EDE1 - HRR25 as ITSN2 - TTBK2, EDE1 - PKC1 as ITSN2 - PKN3, and EDE1 - PRK1 as ITSN2 - AAK1, all of which are mediated by ITSN2's fifth SH3 domain. None of these human ITSN1 or ITSN2 interologs have been identified by traditional PPI mapping physical experimental techniques.[50] Furthermore, their interaction partners do have orthologs in worm, none of which are known to interact with worm ITSN-1, suggesting that there may be a gap in the experimental interactomic data. However, when we use MUSCLE[4] to create multiple sequence alignments of each set of these human and yeast interactors, and their worm orthologs, we find that the specific binding site in the human interactor is not conserved in either the yeast or worm orthologs (see Table 3-6). In our observations, unconserved interactions are often entirely missing the target ligand, which is consistent with earlier work comparing the worm and yeast SH3 interactomes.[46]

Thus, there are many more unconserved than conserved interactions within this protein family. Some of these are due to unmatched interactions – those that are not found in the other species even though an ortholog exists, but many are due to unmatchable interactions – those that have no ortholog to match to. Some lack of conservation may be due to false negative and positive PPIs in either species interactome, but this likely doesn't explain all the lack of conservation as the intersectin gene family has higher interaction conservation than most: with 16, human ITSN1 and worm ITSN-1 have more interologous partners than most worm-human orthologs, ranking in the 99.8th percentile (see Figure 3-9). Additionally, breaking down the interactions to the level of their mediating domains shows that SH3 domains seem to conserve few, if any, interactions, possibly due to rapid evolution of binding motifs and domain architectures as previously found between worm and yeast.[46] While it is often assumed that protein domains are indicative of a protein's function and that a conserved domain architecture indicates a conserved function, the

lack of consistency between the binding partners of orthologous yeast, worm, and human SH3 domains suggests that this phenomenon may be more complicated.

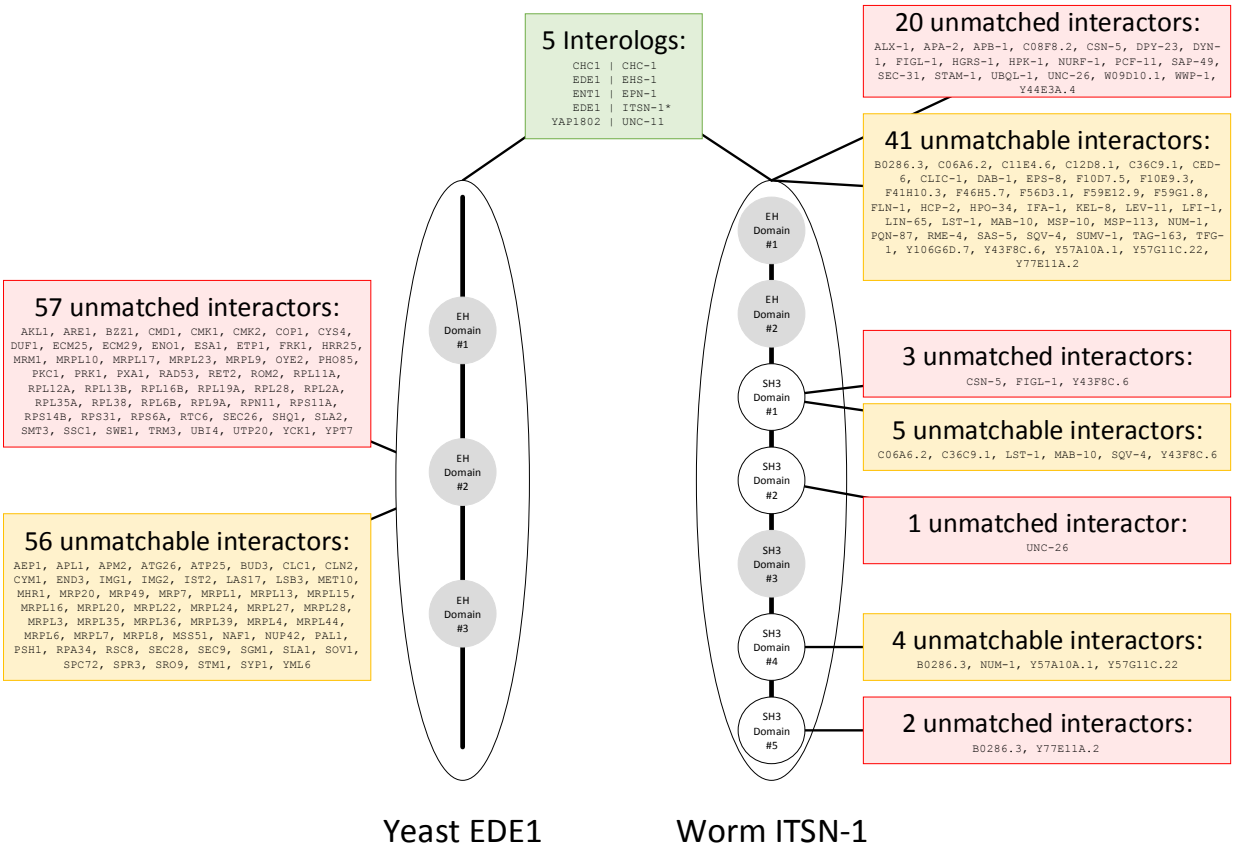


Figure 3-8 – Schematic of interaction conservation between yeast EDE1 and worm ITSN-1. The two proteins share five interologs. All other interactions are divided into unmatchable, if the interaction partner has no ortholog in the other species, and unmatched, if there were no interactions between any of the orthologs of the interaction partner and EDE/ITSN-1. * indicates self-interactions, which are often excluded from standard PPIN analysis.

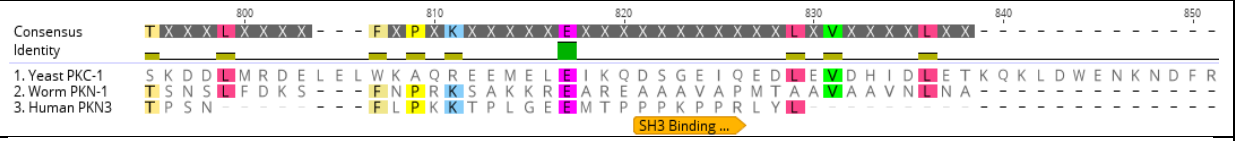
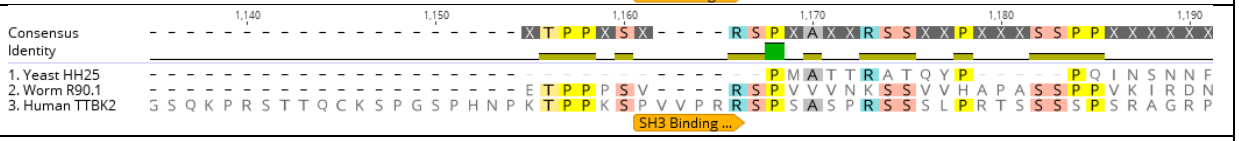
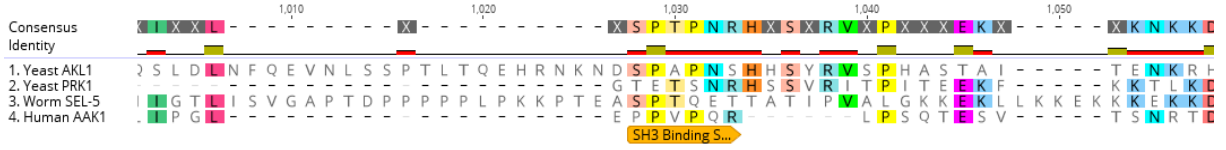
Human ITSN1 Interactor	Worm Interactor Ortholog	Yeast EDE1 Interactor	
PKN3	PKN-1	PKC1	
TTBK2	R90.1	HRR25	
Human ITSN2 Interactor	**Worm Interactor Ortholog**	**Yeast EDE1 Interactor**	
AAK1	SEL-5	AKL1	
AAK1		PRK1	
PKN3	PKN-1	PKC1	Same as with ITSN1
TTBK2	R90.1	HRR25	Same as with ITSN1

Table 3-6 – Multiple sequence alignments of human ITSN1/2 SH3 binding site regions with yeast and worm orthologs. There are 6 PPIs for yeast EDE1 that are conserved and mediated by SH3 domains in human ITSN1/2. We used MUSCLE[4] to perform multiple sequence alignment of the EDE1 and ITSN1/2 interactors, as well as the corresponding worm ortholog, for which there are no known PPIs. Illustrated are the regions around the SH3 binding site in the multiple sequence alignment, with the SH3 binding site in human highlighted in orange. In all cases, the binding site in the human interactors is unconserved in the other orthologs, though two of the bound prolines in human AAK1 are conserved in yeast AKL1. We observed this non-conservation commonly for SH3 binding targets, suggesting that even when PPIs are seemingly conserved, there may be complex evolutionary dynamics at work that lead to expected non-conservation of PPIs. Alignment views were generated using Geneious version 9.[189]

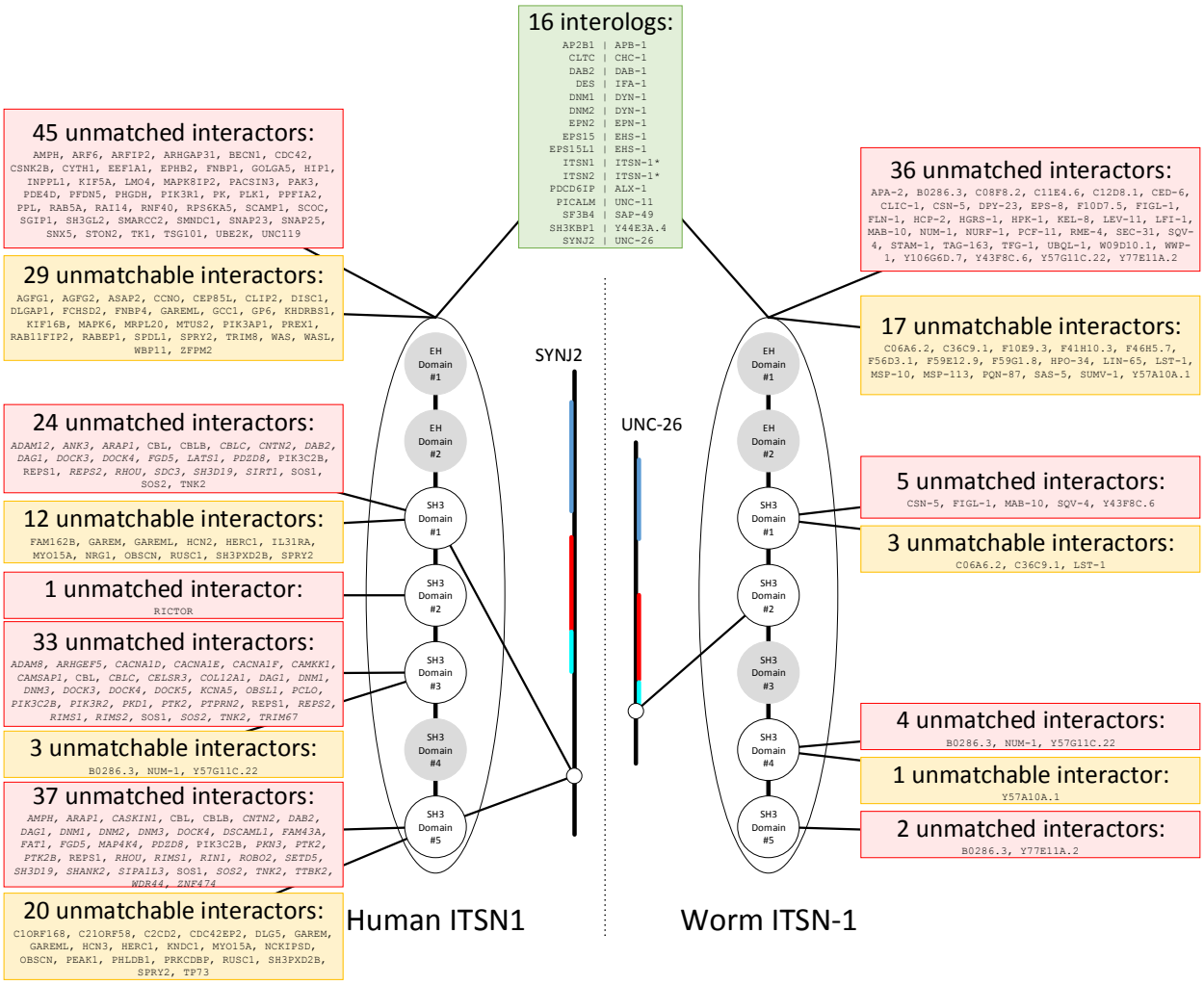


Figure 3-9 – Schematic of interaction conservation between human ITSN1 and worm ITSN-1. The two proteins share 16 interologs; only one with SH3 binding target information available, shown in the middle (human SYNJ2 and worm UNC-26). All other interactions are divided into *unmatchable*, if the interaction partner has no ortholog in the other species, and *unmatched*, if there were no interactions between any of the orthologs of the interaction partner and ITSN1/ITSN-1. The SH3 binding specificity is shown on the two target proteins SYNJ2 and UNC-26, as well as their SAC domains (blue), endonuclease/exonuclease/phosphatase domains (red), and DUF1866 domains (cyan). Human ITSN1 SH3 binding targets are predicted, whereas worm ITSN-1 SH3 binding targets were experimentally determined. Predicted human SH3 interactions without experimental verification are italicized. * indicates self-interactions, which are often excluded from standard PPIN analysis.

3.3 Discussion

We have shown that, based on current data, most PPIs are not conserved and that it is likely there are multiple mechanisms of molecular evolution that rewire PPINs between species. From this, we argue for a model of PPIN evolution that incorporates multiple mechanisms that affect local network regions differently.

3.3.1 Local is better than global for network alignment

The low rate of interolog conservation contradicts a key assumption underlying global network alignment research. Assuming that established ortholog pairs truly represent conserved proteins, a global network alignment algorithm could be expected to align all these orthologs, but it would have to overcome the very high levels of noise generated by unconserved interactions between those orthologs and the misleading signals produced by novel proteins without corresponding orthologs, and possibly cases wherein orthologs have different binding sites. If such an alignment had been optimized on various network-topology-based measures of network alignment quality that correlate strongly with the number of edges aligned, such as edge correctness or symmetric substructure score,[99] the alignment will likely contain many improperly aligned interactions, confusing subsequent efforts to make biologically relevant inferences from the alignment. We propose that network alignment algorithms should be capable of indicating when two networks are and are not (sufficiently) conserved, as is possible with sequence alignment algorithms.

Furthermore, as a result of gene duplication and divergence, interaction conservation rates are not uniform across the interactome. Gene duplications are already a conceptual problem for many network alignment methods[55,73,87] that perform one-to-one protein alignments, as the complexity of the evolutionary relationship cannot be represented in the alignment and the leftover paralogs can then confound the alignment of other proteins. Interaction divergence also indicates that network alignment algorithms should not consider all interactions and areas of the network to be uniformly informative. Instead, a network alignment algorithm seeking to produce an evolutionarily *true* global alignment should also consider gene duplication and divergence, possibly incorporating it explicitly into a scoring function. For example, as Figure 3-5 suggests, PPIN alignment could consider gene family size, scoring larger families lower than smaller families. Alternately, it may be prudent for network alignment algorithms to avoid heavily duplicated proteins, or provide lower confidence scores for alignments of those proteins. More generally, due to the low level of interaction conservation observed in current network data, network alignment methods should not include the number of edges aligned as an optimization or evaluation metric, as many of the aligned edges are likely false and may even distort the proper alignment of nodes.

In our work, we also show that small changes in protein sequence can have large impacts on interaction conservation, in particular small changes in domain and binding site sequences,

which concurs with evidence that small sequence motifs can be used to predict PPIs with high accuracy.[190] More work needs to be done to assess how the sequence similarity and network topology similarity components of most network alignment techniques interact, and how they contribute to algorithmic performance on biologically relevant evaluation measures. Just as sequence alignment is supported by models of gene and protein evolution, network alignment should be supported by models of PPIN evolution, which need to be better developed. We believe that more comprehensive, small-scale analyses of PPINs across species, such as by Kappei et al. 2017,[119] are necessary to develop these models, and serve as a sturdy foundation for more global PPIN evolution research.

3.3.2 PPI Data Quality

One possible explanation for our results is that the protein-protein interaction data is erroneous and incomplete. The PPI data used in this work were retrieved from iRefIndex (see 3.5 Methods), which aggregates interaction data from many PPI databases.[10] False positive interactions could result in systematic errors in our analyses even if there were no systematic bias in how they were generated. False positive interactions are likely to decrease the perceived rate of interaction conservation, as they cannot, by definition, be conserved and are unlikely to randomly match up with another, true positive interaction due to the assumed sparsity of PPINs. False negative interactions may also decrease the perceived rate of interaction conservation, if the true rates of conservation and false negatives are both high, as conserved interactions are erroneously left out of PPI datasets. Consequently, our work may underestimate how well conserved PPINs truly are, though other work has also suggested that this rate is lower than commonly assumed.[46,105]

To consider these data quality issues, we re-performed the analyses in the first two sections using limited subsets of proteins: first with only disease-associated proteins as annotated by OMIM's Morbid Map,[191] and also with only "core" conserved proteins with orthologs in all five species, under the assumption that the proteins in these subsets are relatively well-studied. Interaction conservation numbers were slightly increased, which was expected given the removal of poorly conserved, poorly studied proteins from the data set, while the relationship between domain and interaction numbers remained approximately the same. Thus, our major conclusions are not affected by bias towards the most studied proteins.

Other technical issues include that the unknown number of false negatives means that there is no reliable baseline to use to normalize the differing levels of coverage for the PPI datasets of different species or sections of an interactome and also that it is difficult to prove an interaction can never occur. Despite these various technical issues, it remains useful to evaluate existing theories of network evolution in the context of the entire interactome, using the most comprehensive data available. However, we recommend that experimental studies focus on collecting complete (positive and negative) information for tractable subsets of the interactome, such as individual genes or pathways, across species to inform network evolution studies.

3.3.3 Quality of binding site data

Computational identification of protein domains, as performed by Pfam, is based on matching protein peptide sequences to hidden Markov models that represent specific domain profiles.[192] Due to the costs associated with creating crystal-based structures for proteins, these domain profiles themselves are generated based on previously identified canonical domains, which have typically been well-studied and have known crystal structures. However, many protein domains are known only through electronic annotation processes like Pfam's. Consequently, the set of "known" SH3 domains is likely not unbiased, with a skew towards the canonical SH3 domains that serve as the core blueprint for the definitive HMMs, which may have undue effects on any analyses of SH3 domain evolution.

Furthermore, though SH3 domains are relatively well-studied, partially due to their binding affinity for a well-defined short, linear motif, we continue to learn new facts about the SH3 domain. Recent work with human SH3 domains identified, based on their peptide-binding preferences, seven additional classes of SH3 domains, in addition to the two canonical classes.[50]

3.4 Conclusion

Based on the latest available PPI data, there is little evidence to support the idea that PPI conservation is high enough to support analytical techniques such as interaction transfer between species or global network alignment. Considering evolutionary concepts such as duplication and divergence of genes and the presence of interaction-mediating domains, further inconsistencies arise with the presumption of high, uniform rates of interaction conservation between species. Whether these inconsistencies arise due to data quality and completeness issues or not, however,

is difficult to determine under the current circumstances, and a radical revamp of how PPI data is collected may be required to fully unravel the problem.

Meanwhile, these and other issues with PPI data should prompt more work in understanding the fundamentals of interactome evolution, instead of the development of high-level computational methods that may be ignoring, or even obscuring, the biological mechanisms at play. We believe it'd be more prudent, for example, for researchers to focus more on local network alignment rather than global network alignment so as to establish what biologically relevant information PPIN alignments can and should extract from interactome data, or to investigate what factors suggest that an interaction can be mapped from one species. These are critical questions, needed to establish the baseline utility and reliability of inter-species PPIN data analysis, that should be answered before the competitive development of algorithms to solve artificial questions that may not even reflect true biological reality.

3.5 Methods

3.5.1 Protein-protein Interaction Data

Ensembl version 89 served as the authoritative source for gene and protein data, from human (*H.sapiens*), mouse (*M. musculus*), fruitfly (*D. melanogaster*), worm (*C. elegans*) and yeast (*S. cerevisiae*).[193]

Species	Proteins	Interactions	Mean Interactions per Protein	% IRefIndex Physical Interactions Captured
S. cerevisiae	5 755	82 137	14.272	99.76
C. elegans	4 780	12 040	2.519	97.71
D. melanogaster	8 280	39 430	4.762	95.68
M. musculus	6 082	15 998	2.630	97.99
H. sapiens	14 286	178 337	12.483	99.06

Table 3-7 – Summary of interaction data used.

The protein-protein interaction networks in this work were compiled based on protein-protein interaction data retrieved from iRefIndex version 14.0.[10] This data uses controlled vocabulary codes from the Molecular Interactions Controlled Vocabulary for only intra-species, physical interactions (MI:0218 and MI:0915), which excludes predicted interactions (MI:1110). Interactions identified only via inference, interaction prediction, or an unknown method

(MI:0362, MI:0063, and MI:0686) were also filtered out. All interactions were mapped to the corresponding Ensembl proteins by HGNC symbol, UniProtKB ID, Entrez Gene ID, and/or RefSeq ID, with a success rate of no less than 95% (95.68% for *D. melanogaster*).[194-197] Complexes were resolved as multiple two-way protein interactions between all participating proteins. Self-interactions were included as self-loops in the networks to consider PPIs created by gene fusion and gene fission events. Proteins with no interacting partners were excluded from the data set. When multiple proteins had identical sequences, the duplicates were removed in favour of the first alphabetically (arbitrary selection) so as to prevent skewing of the data analysis.

Certain highly promiscuous or unusually long proteins, and their orthologs, were removed from later analyses as outliers to prevent distortion of results:

- human EED (460 interactions, 466 amino acids), mouse EED (1140 interactions, 441 amino acids), fruitfly ESC (13 interactions, 425 amino acids), fruitfly ESCL (2 interactions, 462 amino acids), & worm MES-6 (6 interactions, 459 amino acids)
- human OBSCN (9 interactions, 8925 amino acids) & mouse OBSCN (2 interactions, 8032 amino acids)
- human TTN (107 interactions, 35991 amino acids), mouse TTN (24 interactions, 35213 amino acids), & worm TTN-1 (1 interaction, 18562 amino acids)
- human UBC (9812 interactions, 685 amino acids), mouse UBC (324 interactions, 734 amino acids), and yeast UBI4 (3192 interactions, 381 amino acids).

3.5.2 Domain Data

Domain data for all proteins was retrieved from Ensembl. For identification of SH3 domains, we began with a union of the sequences identified by Pfam, PROSITE, SMART, and Superfamily,[192,198-200] and retrieved all domains with InterPro terms IPR001452 (SH3 domain), IPR003646 (SH3-like domain, bacterial-type), and IPR011511 (Variant SH3 domain).[201] We then manually assessed how likely each sequence was a true SH3 domain based on concurrence between the databases, key sequence signatures such as the canonical GXXP sequence, and corroboration in the literature.

From these assessments, Pfam was found to be the most reliable of the databases, and was then used as the database for identification of other domains. Protein-binding domains were identified using InterPro terms tagged with GO annotation 0005515 (protein binding).[92]

The peptide recognition modules used in this work were initially curated based on work by Castagnoli et al..[202] Of the 16 PRMs listed, 7 were removed because they were not tagged with the protein-binding GO annotation in Interpro or because their Interpro description indicated they perform a function other than protein-binding. The remaining 9 PRMs are: Bromo, FHA, GYF, PDZ, Polo-box, PTB, SH2, SH3, and WW. FHA, GYF, Polo-box, and PTB domains were excluded from Table 3-3 due to low counts, but were included in the overall PRM counts.

Species	# SH3 Domains	# Proteins w/ SH3 Domains	# Human Orthologs to SH3 Proteins	# Orthologs to Human SH3 proteins
S. cerevisiae	27	23	35	28
C. elegans	80	60	139	69
D. melanogaster	84	64	165	79
M. musculus	285	210	240	227
H. sapiens	291	214	-	-

Table 3-8 – Summary of SH3 complements for several species. Note that each species has more SH3 domains than SH3 proteins, as many SH3 proteins have multiple SH3 domains. Note further that due to gene duplication, many human SH3 proteins share the same orthology, i.e. are paralogs of each other.

Orthology data was compiled using a union of Ensembl, InParanoid, and OrthoMCL.[193,203,204]

3.5.3 SH3 Domain Binding Data

Experimental data on the binding targets of yeast and worm SH3 domains was compiled from previous work yeast and worm.[46,49] and human SH3 domains. Yeast and worm SH3 domain binding affinities were experimentally determined using phage display on individual domain isolates, and then confirmed using yeast-two-hybrid experiments on whole proteins.

Human SH3 data were predicted using DoMoPred[188] based on phage display data from [50], with interactions included only if predicted with 95% confidence or higher.

Chapter 4

Small-Scale Visualization of Protein-Protein Interaction Network Alignment

I conceived of, designed, and conducted this project. Gary D. Bader supervised and advised this project.

4 Small-Scale Visualization of Protein-Protein Interaction Network Alignment

4.1 Introduction

In interactomics at large, there remain many key outstanding questions to be answered. At what rate are PPIs conserved? Is this rate substantive and consistent enough to support various inter-species interactomic methods such as network alignment? What factors control the rate of PPI conservation? The answers to these questions are critical to our understanding of biological system evolution, and the development and refinement of bioinformatic tools to derive insights from interactomic data.

For network alignment, a major concern is to identify which proteins should be aligned and which proteins should not (see 1.1 Introduction). In the current state of the field, global network alignment methods presume that entire interactomes should be aligned to each other, which is likely unfounded given gene gains and losses between species. Furthermore, network alignment quality is measured with two types of measures, typically averaged across the entire alignment. The first are abstract network topological measures that may or may not capture any biologically relevant phenomena, created and used almost exclusively within the network alignment community. The second are gene functional measures, such as the ability to predict GO functional terms, but state-of-the-art network aligners perform quite poorly on these measures, especially considering that many utilize sequence alignment as input data, which is expected to strongly correlate with gene function.

Similarly, interolog mapping is a technique used in computational biology that may or may not be generally appropriate. While originally recommended only for proteins that were best reciprocal BLASTP hits,[178] others have used interolog mapping more liberally and explored where the limits should be set.[21,104,166] Like with network alignment, these efforts have been applied to the interactome altogether, disregarding the possibility that the interactome may not be uniformly conserved. Paralogs, in particular, may pose a sticking point: paralogs have lower rates of interaction conservation than expected (see 3.2.1 Interolog conservation across species) and only one paralog can be a reciprocal best BLAST hit for an ortholog, suggesting that extra care must be taken with paralogs when performing interaction mapping.

For researchers interested in using network alignment, interlog mapping, and other network-based methods in their own research, these questions of suitability are difficult to answer. Interactomics is a "big data" field; yet molecular biology is often conducted on a small scale, researchers investigating one protein, or one pathway, or one complex, or one biological function at a time. With "big data" interactomic methods making sweeping, possibly inaccurate, generalizations across the interactome, researchers interested in incorporating interactomics into their own smaller-scale research can be left unsure of how to do so, or whether it would even be appropriate to do so. Attempting to incorporate interactomic data and methods in their own work can be quite difficult, due to the dearth of tools to bridge the gap between "big data" interactomics and "small data" studies, such as that done by Kappei et al..[119]

To address this, we propose a new framework for utilizing network alignment, one focused on facilitating the generation and evaluation of specific, smaller in scope hypotheses. We have also developed a prototype small-scale network visualizer, the Pairwise Protein Alignment Analysis Tool (PPAAT), designed to present pairs of proteins and their immediate interactomic neighbourhood, integrated with other data sources.

4.2 Results

4.2.1 Description of PPAAT

PPAAT is an HTML/JavaScript application written using Cytoscape.js designed to visualize two proteins and their neighbourhoods at a time. PPAAT is specifically designed to visualize only two proteins at once to allow for data integration and more details to appear on the screen. When the user selects two proteins from different species, PPAAT loads protein-protein interaction network data, domain data, and ortholog data for display (see Figure 4-1).

The user's two query proteins are displayed prominently near the centre of the display, with their sequence stretched vertically from the N-terminus at the top to the C-terminus at the bottom. Along the query proteins' sequences, we display the peptide recognition modules on that protein, specifically BROMO, EH, FHA, GYF, PDZ, Polo, PTB, SH3, SH3, and WW domains, being domains whose primary function is protein-binding.[192,201,202] The domains are arranged in protein sequence order, spaced evenly for visual clarity.

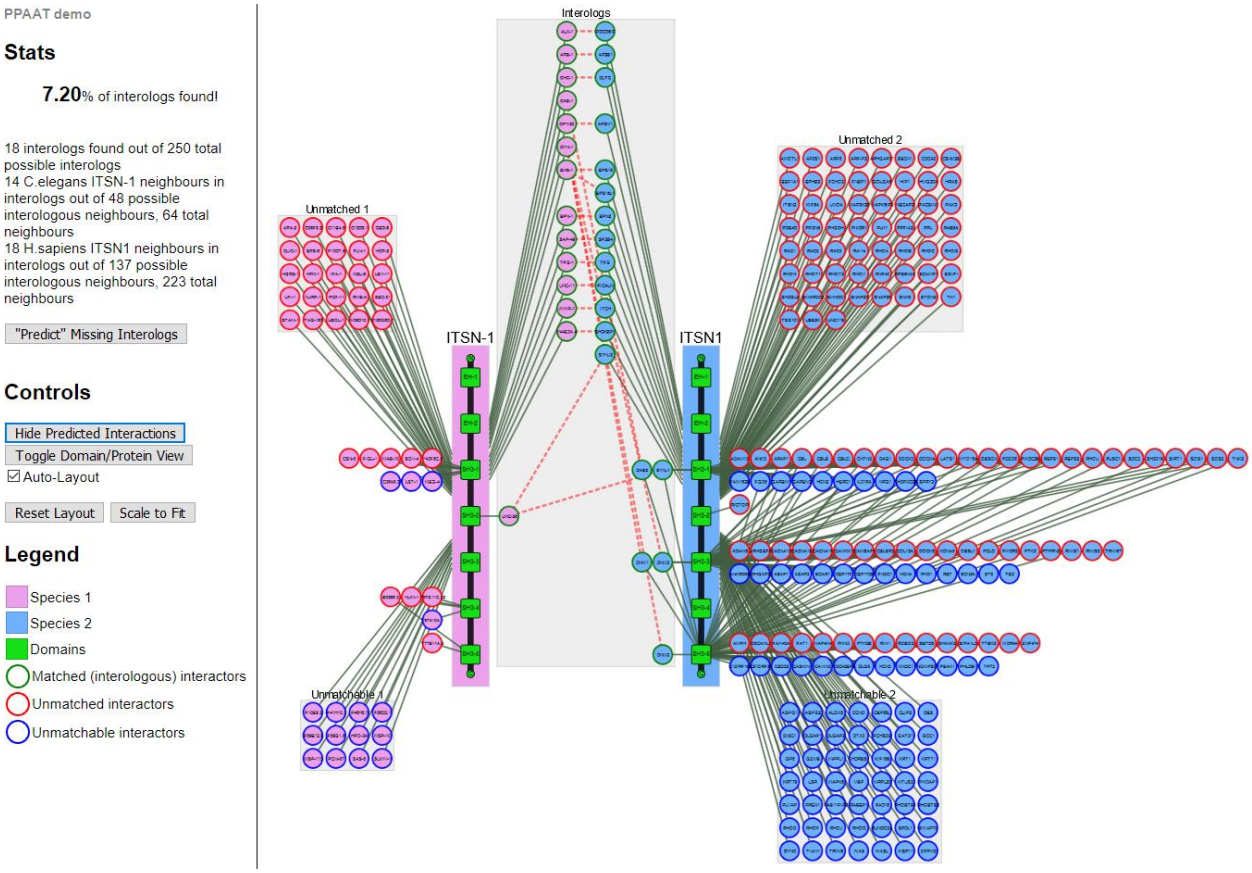


Figure 4-1 – The default PPAAT view. On the left is *C. elegans* protein ITSN-1 and its interactors, in pink. On the right is the *H. sapiens* ortholog ITSN1 and its interactors, in blue. These two proteins have a particularly high rate of interologs between their neighbours, indicative of their orthology. Their protein domains, in green, are arranged vertically along the length of each protein. Matched, interologous interactor nodes are encircled in green, unmatched interactor nodes are encircled in red, and unmatchable interactor nodes are encircled in blue. Node categories are grouped where needed for clarity of visual presentation.

Using the network data, all the neighbours of both query proteins are loaded and displayed. Neighbours are grouped and displayed according to two separate priority systems. Firstly, when information is available, we connect neighbours directly to the specific domain in the respective query protein; otherwise, the neighbour is connected to the query protein itself. If there are independent data indicating that the neighbour interacts with the domain and with the protein, we prioritize the domain-neighbour interaction to the exclusion of the query-neighbour interaction, presuming that the domain is responsible for facilitating the interaction. If more than one domain facilitates the interaction, we arbitrarily group the neighbour with the first of the interacting domains.

We also group neighbours based on their ortholog in the opposing query species. If a neighbour has no orthologs in the opposing species, then that neighbour is considered “unmatchable,” as it could not possibly form an interolog with the query proteins. If a neighbour has at least one ortholog in the opposing species, and at least one of those orthologs interacts with the opposing query protein, then the neighbour, its ortholog, and the two query proteins form an interolog. Finally, if the neighbour has orthologs but those orthologs do not interact with the opposing query protein, then that neighbour is considered “unmatched.”

By grouping interacting proteins into these three categories, the percentage of interologs shared for the two query proteins can be accurately calculated as the number of matched neighbours divided by the number of unmatched neighbours. This value is computed and displayed on the left panel for users, as well as a listing of the interologous interactions that are “missing.” There are also options for the user to change whether predicted PPIs should be displayed and whether the domains should be displayed (see Figure 4-2). All nodes in the visualization also have a dropdown menu accessible via right-mouse-click that provides additional information and external links for each protein (see Figure 4-3).

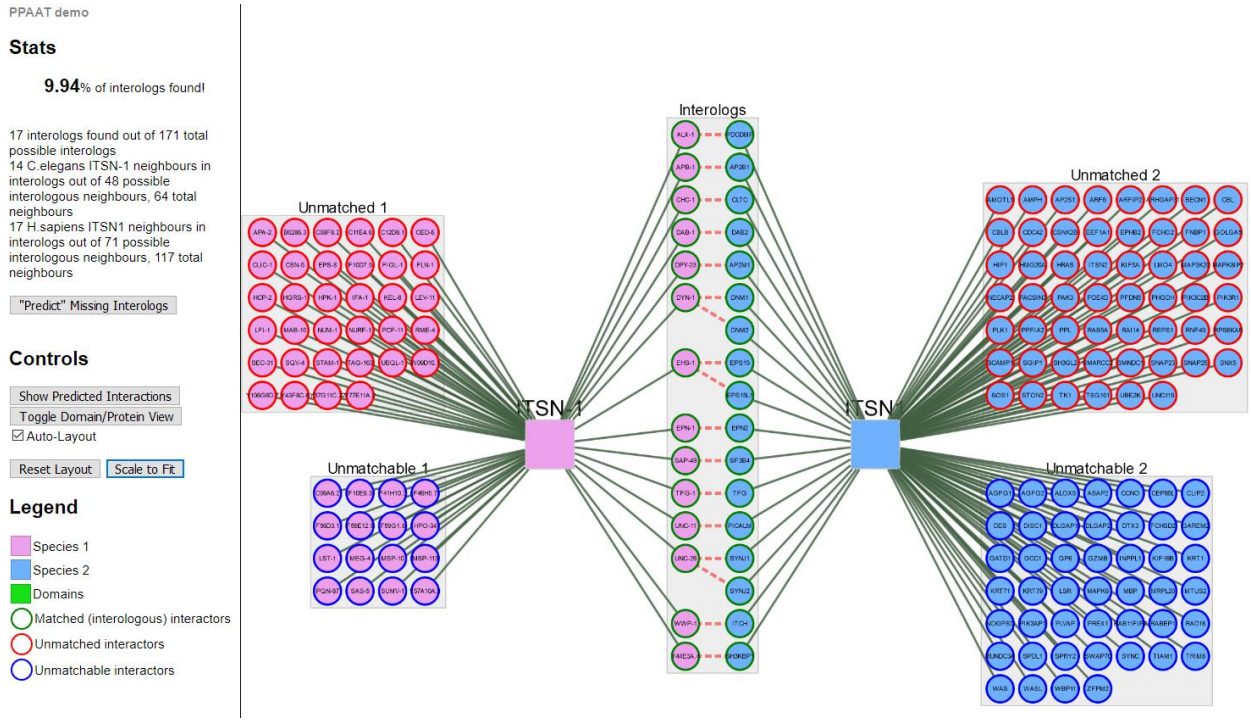


Figure 4-2 – A reduced PPAAT view. PPAAT can hide visual features undesired by the user, such as predicted interactions and the domains of the query proteins. On the left is *C. elegans* protein ITSN-1 and its interactors, in pink, and on the right is the *H. sapiens* ortholog ITSN1 and its interactors, in blue, as in Figure **4-1**.

PPAAT relies on data files manually downloaded from Ensembl,[156] Inparanoid,[157] OrthoMCL,[158] and BioGRID[9] or iRefIndex.[10] These files are processed into smaller intermediate files, extracting cellular, experimentally identified PPIs, including from protein complexes. These files then parsed into a JSON file for PPAAT display by a series of Java scripts, taking approximately 10 seconds on a 3.4 GHz processor, including file IO. We expect that the runtime could be significantly decreased with refactoring and more efficient preprocessing, including using an online database rather than an offline filesystem.

Figure 4-3 – Dropdown menu for proteins in PPAAT. PPAAT includes additional information and links in dropdown menus for each protein, accessible via right-mouse-click. These menus allow users to access more information about the protein and functionality that could not otherwise be displayed in the larger view.

4.2.2 Use Cases

While PPAAT is still in early development, it supports a variety of use cases pertinent to molecular biology research.

The user has two putative orthologs and would like to evaluate whether PPI network data supports the orthology hypothesis.

> Given conservation of protein-protein interactions, it may be expected that orthologs should share a higher proportion of orthologous neighbours, or interologs, than non-orthologous proteins. Some studies have indicated that interaction and interolog conservation rates are quite low, especially when the proteins have paralogs, but some

ortholog pairs may yet have a higher rate of interaction conservation that would make them observable statistical outliers (see Figure 4-1 and Figure 4-4). By identifying when proteins have unusually high interolog ratios, PPI data might be used as another evidence source for ortholog identification and to evaluate the potential level of conservation of gene function.

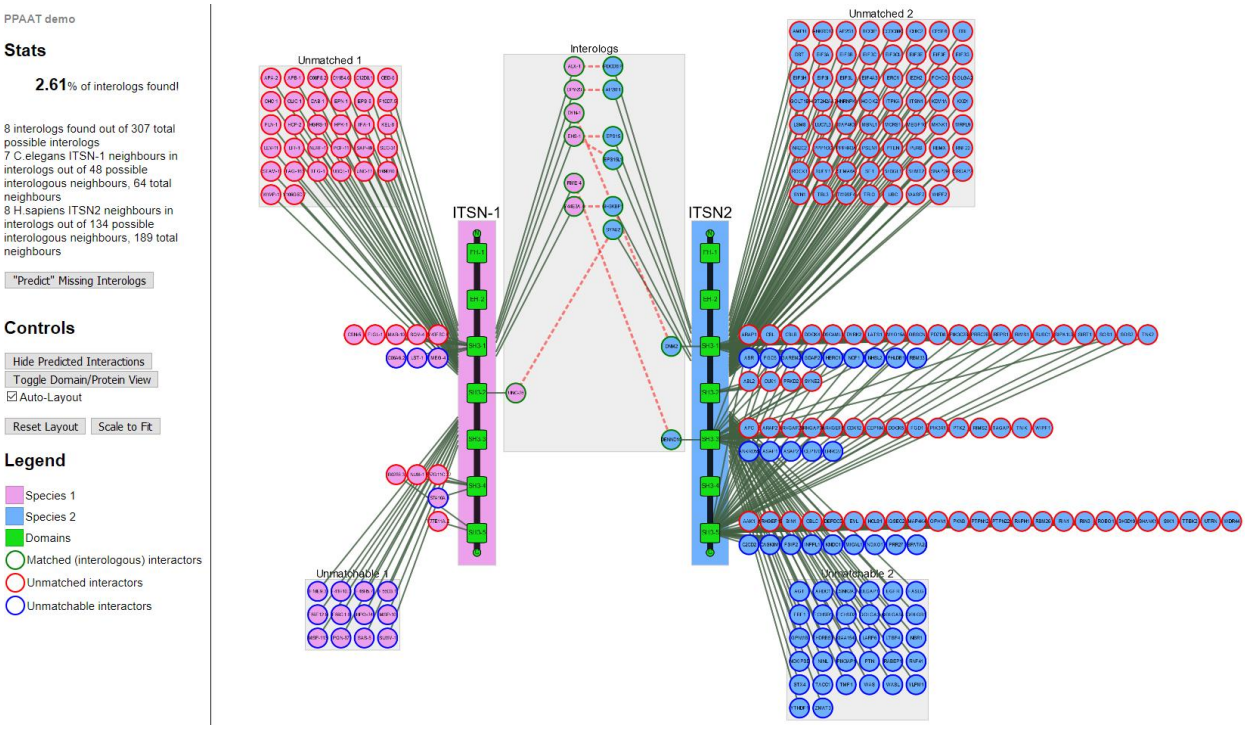


Figure 4-4 – A PPAAT visualization of paralogs *C. elegans* ITSN-1 and *H. sapiens* ITSN2. On the left is *C. elegans* protein ITSN-1 and its interactors, in pink. On the right is *H. sapiens* paralog ITSN2 and its interactors, in blue. In contrast to Figure **4-1**, while there are some interologs matched between the two proteins, there are substantially fewer than between the two orthologous proteins. However, without contextual information, it is unclear whether a 2.61% interolog rate is significant or not.

By also considering the domain data provided by PPAAT, the user may also consider network rewiring in the context of sequence changes within the two proteins. A gained/lost domain may account for some level of rewiring between the two proteins, explaining a lower than expected rate of interaction conservation. Conversely, major changes in domain complement may also disqualify putative orthologs (see Figure 4-5).

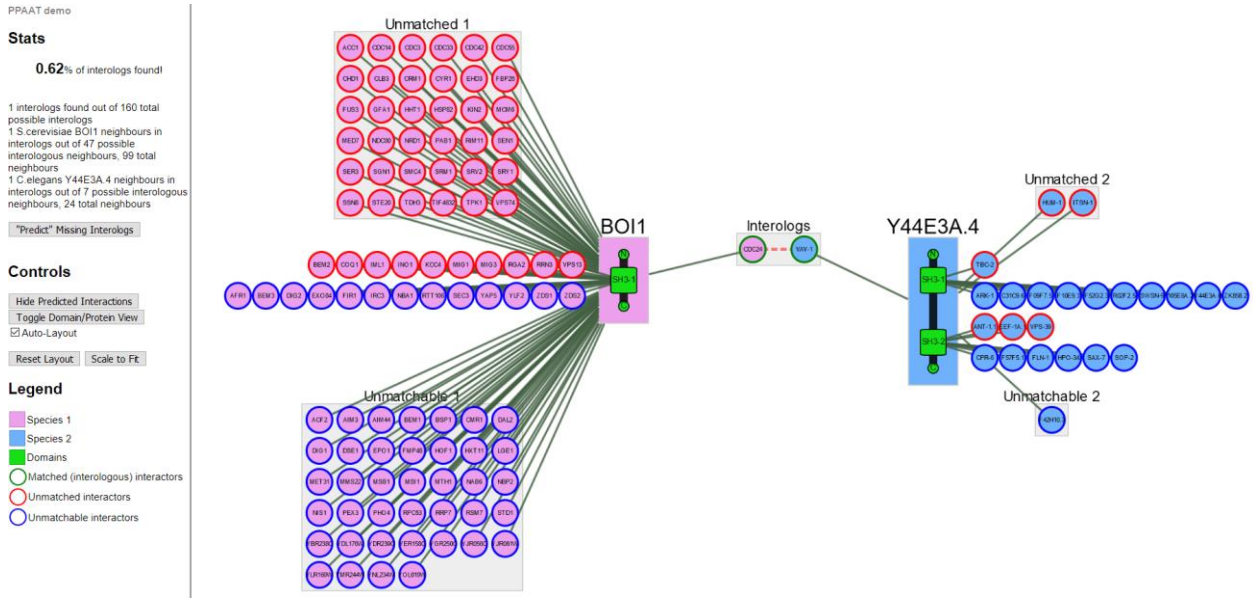


Figure 4-5 – A PPAAT visualization of non-homologous proteins *S. cerevisiae* BOI1 and *C. elegans* Y44E3A.4. On the left is *S. cerevisiae* BOI1 and its interactors, in pink. On the right is *C. elegans* Y44E3A.4 and its interactors, in blue. In contrast to Figure **4-1** and Figure **4-4**, the interolog matched rate is very low and the domain architectures of the two proteins are different, suggesting correctly indicating that the two proteins are not homologous.

The user has two orthologs and would like to evaluate the completeness of their interaction data, or identify possible false negative protein-protein interactions involving either ortholog.

If a user expects a certain interaction conservation rate between the two proteins of interest, PPAAT could be used to evaluate whether the extant PPI data meets this expectation. In the event that the observed interaction rate is much lower than expected, one may conclude that there are interactions missing. In this context, PPAAT can be considered a PPI predictor, using interolog completion as its predictive indicator. PPAAT provides users a list of these predicted missing PPIs for both the two input proteins (see Figure 4-6). As one missing PPI could complete multiple interologs, arising from duplication of one of the interacting genes, missing PPIs that would complete more interologs are listed first, reflecting the relatively stronger strength of these predictions. These interaction predictions can be targeted for further analysis or for PPI mapping experiments, to confirm their absences from the interactome.

Predicted C.elegans ITSN-1 interactions		
# Missed Interologs	Predicted Interactor	Indicating ITSN1 Interactors
4	CED-10	RAC1, RAC2, RAC3, RHOG
	RAC-2	RAC1, RAC2, RAC3, RHOG
3	CED-5	DOCK3, DOCK4, DOCK5
	RHO-1	RHOA, RHOB, RHOC
	SLI-1	CBL, CBLB, CBLC
2	AAP-1	PIK3R1, PIK3R2
	ADM-2	ADAM12, ADAM8
	AEX-4	SNAP23, SNAP25
	CHW-1	RHOU, RHOV
	EGL-19	CACNA1D, CACNA1F
	FCHO-1	FCHO2, SGIP1
	HIM-4	OBSCN, OBSL1
	KIN-32	PTK2, PTK2B
	MIRO-1	RHOT1, RHOT2
	MIRO-2	RHOT1, RHOT2
	REPS-1	REPS1, REPS2
	RIC-4	SNAP23, SNAP25
	SOS-1	SOS1, SOS2
	UNC-10	RIMS1, RIMS2
	UNC-52	OBSCN, OBSL1
	UNC-89	OBSCN, OBSL1
1	AMPH-1	AMPH
	APS-2	AP2S1
	ARF-6	ARF6
	B0303.7	SH3D19
	BEC-1	BECN1
	C25H3.1	TTBK2
	C28G1.10	UBE2K
	C31C9.2	PHGDH
	C34B2.4	LMO4
	C53B4.4	PDZD8
	CDC-42	CDC42
	CDH-4	FAT1
	CKK-1	CAMKK1

Predicted H.sapiens ITSN1 interactions		
# Missed Interologs	Predicted Interactor	Indicating ITSN-1 Interactors
1	ANKS1A	C11E4.6
	ANKS1B	C11E4.6
	AP1B1	APB-1
	AP2A1	APA-2
	AP2A2	APA-2
	BPTF	NURF-1
	C17ORF49	Y43F8C.6
	CAGE1	HCP-2
	CD2AP	Y44E3A.4
	CEP250	LFI-1
	CLTA	CLIC-1
	CLTB	CLIC-1
	CLTCL1	CHC-1
	COPS5	CSN-5
	CROCC	LFI-1
	CROCC2	LFI-1
	DAB1	DAB-1
	DENND1A	RME-4
	DENND1B	RME-4
	DENND1C	RME-4
	EPN1	EPN-1
	EPN3	EPN-1
	EPS8	EPS-8
	EPS8L1	EPS-8
	EPS8L2	EPS-8
	EPS8L3	EPS-8
	FIGN	FIGL-1
	FIGNL1	FIGL-1
	FIGNL2	FIGL-1
	FLNA	FLN-1
	FLNB	FLN-1
	FLNC	FLN-1
	FUBP1	C12D8.1
	FUBP3	C12D8.1

Figure 4-6 – PPAAT Predicted PPIs for *C. elegans* ITSN-1 and *H. sapiens* ITSN1 based on interolog conservation. PPAT can predict missing PPIs presuming interolog conservation. PPIs are predicted separately for each input protein, with the predicting orthologous interactors listed as evidence for closer examination. As some missing PPIs could satisfy multiple interologs if found, PPAAT considers these stronger PPI predictions and prioritizes them in the list order.

The additional domain-specific interaction data presented in PPAAT may also help simplify this work for the user. If a missing interaction would be paired with a domain-mediated interaction on the other protein, the user may want to specifically seek a domain-mediated interaction, rather than a generic PPI. With well-studied domains, this could be a much simpler task, using computational models for domain binding specificity.[205]

The user has two orthologs and would like to identify additional ortholog pairs from their neighbours.

Alternately, if a user finds that the interaction conservation identified for the two input proteins is lower than expected, they may instead attribute this discrepancy to missing

orthology pairings between neighbours of the input proteins. By examining the proteins listed as “unmatched” for both input proteins, the user may find previously unidentified ortholog pairs, which would then increase the number of interologs observed.

The domain data provided by PPAAT can also be used in this process. With the unmatched neighbours specifically attributed to domains on either protein, the user can specifically focus on the matching domain on the other protein, if such a domain exists, and its unmatched interactors.

The user has two orthologs and would like to evaluate the interaction conservation between them.

Some users may not have prior beliefs about interaction conservation rate, and are instead seeking to clarify their understanding of it. PPAAT offers a tool for users to spot-check interaction conservation at various locations across the interactome. This may be particularly useful if the user believes that interaction conservation rates vary across the interactome. If the user is studying a particular set of orthologs they believe may have an unusual pattern of interaction conservation, such as a well-conserved pathway or proteins with a particular function, PPAAT allows a user to quickly validate this belief without writing their own analysis scripts.

The user has a network alignment and would like to assess the correctness of this alignment, with respect to a given protein.

Network aligners often produce very large alignments that are optimized based on multiple features and across many proteins. For researchers interested in a single protein or a small set of proteins, this can be an overwhelming amount of data, and the alignment may have ultimately misaligned their proteins of interest in any case. PPAAT offers a way for users to quickly pare down an alignment to just their protein of interest and its aligned partner, then verify whether this specific alignment was justifiable or not based on the protein’s domain signature and their surrounding network neighbourhood. For this use case, PPAAT could be used as a viewer app for a network alignment tool.

The user has two orthologs and would like to evaluate the appropriateness of interolog mapping from one to the other.

> Interolog mapping between orthologs is commonly performed to transfer PPI data from a well-mapped species to a poorly-mapped species. However, the general advisability of the method is debatable, given low interaction conservation rates. Nevertheless, PPAAT could be used to transfer interactions from one ortholog to another, using its listings of unmatched neighbours. The domain information provided by PPAAT can be a safeguard, disqualifying certain mappings if, for example, the domain that mediates an interaction-to-be-transferred is absent from the ortholog.

4.3 Discussion

4.3.1 Additional Features to be Implemented

While PPAAT can already be used to answer a variety of biological questions, as described above, it remains an early prototype whose feature-set could yet still be expanded. The integration of more types of data into PPAAT, either explicitly into the application or via connections to other online bioinformatics tools, opens up multiple interesting possibilities.

One major type of data that PPAAT could additionally integrate is functional data. Protein function prediction is often cited as one of the objectives of network alignment; global network alignment seeks to align "functionally orthologous" proteins across the interactome, whereas local network alignment seeks to align orthologous pathways or complexes, which will typically fulfill the same biological function. Function is also more broadly associated with PPIs in general, under the "guilt by association" paradigm, used to predict protein function from their interacting neighbours.[28]

By incorporating protein function as an integrated data source, PPAAT could allow users to reconsider their presumptions about the relationship between function and PPI conservation. With PPAAT organizing the unmatched and unmatchable proteins by their function, protein function could be used to identify where orthology relationships or PPIs are missing from the input proteins' neighbourhoods. A user may find, for example, that PPIs involved in a certain function are more or less conserved than average, which could be a novel pattern of interest to the user and to the wider interactomics community. PPAAT may also reveal that that one of the

input orthologs has neighbours of a unique function, possibly indicating functional divergence, which could influence whether functional transfer should be performed between the two orthologs.

PPAAT's integration of protein sequence data could also be improved. Currently, PPAAT only displays the presence of peptide recognition modules on the two input proteins, evenly spaced for visual clarity. This could be expanded to include more protein domains, so users can more accurately judge the similarity of the two proteins. These domains could also be displayed by their relative positions on the protein, so that in the case of domain gain or loss, when there are multiples of a given domain type, users can quickly determine which of the domains are conserved.

Further considerations include pre-generating network alignments using popular alignment methods and storing their pairings for easy user querying, changing the codebase to allow for visualization of paralogs for visualization of post-duplication network conservation dynamics, and adding filters for the experimental mapping technologies used to detect interactions.

4.3.2 The Broader PPAAT Framework: Rethinking Network Alignment

PPAAT promotes the idea of increasing the value of interactomic data by considering it in small, accessible portions and visually integrating detailed, complementary information to aid interpretation. Currently, while interactomic data in plentiful in various online databases, such as BioGRID and iRefIndex, using the data effectively requires parsing plaintext MITAB files. There are also few tools linking interactomic data to other biological datasets, making interpretation of interactomic data a complex, technically-involved task. Multiple types of tools following this approach could be developed, such as ones that focus on pathways or complexes, protein families, or small network regions, perhaps in isolation or in comparison. Rather than display domain information and organizing around orthologs and interologs, other tools may focus on integrating other data, such as sequence similarity, protein structure, protein function, gene expression, subcellular localization, to facilitate integration of interactomic data into other studies with diverse goals.

We believe that such work is essential for the long-term viability of interactomics. Ultimately, interactomic research is only as useful as its ability to contribute to the understanding of biology,

and interactomic currently stands afield from other molecular biology research. Popular “big data” approaches to interactomic research do not generate easily testable hypotheses and are ill-suited for generating the specific insights sought by researchers, impeding the progress of interactomics. As a result, competing interactomic paradigms can only be contrasted at high levels of abstraction. Instead, low-throughput experiments are needed to develop a complete understanding of the molecular and evolutionary mechanisms that produce the phenomena found in “big data” studies.

While we do not believe that PPAAT alone fill this role, we hope that PPAAT and similar tools facilitate this important work, making interactomics more accessible to non-experts using a “small data” approach. Hopefully, this increased accessibility will convince researchers of the value of interactomics, who will in turn contribute to interactomics by providing new insights, perspectives, and directions for research.

4.4 Conclusion

The Pairwise Protein Alignment Analysis Tool is a novel visualization tool for network alignment. By having a small scope, two query proteins, PPAAT can integrate and visualize different data types to provide users a broad perspective on their proteins of interest. As a minimalistic web-based tool with no workflow integration, PPAAT is flexible enough for easy incorporation into different analyses, making network alignment data and network evolution concepts accessible to biologists.

Analysis tools like PPAAT can empower scientists to draw conclusions, formulate new hypotheses and evaluate network alignment by their own standards. This will in turn inform the network alignment community how to improve their alignment methods, by generating user feedback from the application of network alignment in their own work. Ensuring such tools are open to all network alignment methods will allow members the scientific community to assess which methods are most useful for their purposes, effectively “open-sourcing” the network alignment evaluation process, using practical, realistic criteria. Ultimately, while development of tools like PPAAT will require additional effort from the network alignment community, it will enable further progress and innovation in network alignment research, as well as supporting efforts in related areas of systems biology.

Chapter 5

Summary and Future Directions

5 Summary and Future Directions

5.1 Summary of Thesis

In this thesis, I focus on the extraction of meaningful biological insights from alignment of protein-protein interaction networks (PPINs). I investigate how PPIN alignments correspond with existing biological data on protein-binding domains, interologs, and paralogs, and find that these are difficult elements for current PPIN alignment methods to account for properly. Given these failings, PPIN alignment needs to be reconsidered and reimagined, so that alignment methods can produce alignments that are consistent with this known data. I propose that, in order to do so, PPIN alignment needs to be critically assessed in a small-scale, detailed manner, to develop our understanding of the fundamental mechanics of biological network evolution upon which we can later base high-level alignment and analysis methods.

In Chapter 2, I presented GreedyPlus, an algorithm designed for the alignment of interface-interaction networks (IINs). With newly available IIN data for SH3 domains in *S. cerevisiae* and *C. elegans*, I used GreedyPlus to probe some of the mechanisms that underlaid the popular PPIN alignment algorithms of the time. Keeping GreedyPlus a relatively simple algorithm, I tested how various similarity features impacted alignment quality, and how biological and network topology features should be weighted in network alignment. With proper parameterization, GreedyPlus produced a near ideal alignment of the two networks, capturing all 16 orthologous protein pairs in its alignment, and performed admirably on several simulated yeast SH3 IINs.

After publishing my GreedyPlus work, my initial goal was to iterate on the algorithm, using new *H. sapiens* SH3 binding data or whole PPINs. However, I was troubled by the existence of many similarly performing sets of features and parameters and the prominence of protein BLAST similarity among them, despite my efforts to manage overfitting. Additionally, I was perplexed that various domain-specific features I had experimented with failed to contribute to good alignments. Chapter 3 consists of the fruitful results of various attempts to investigate these issues from alternate angles, such as by tracing SH3 domain ancestry, considering paralogs and interologs, and examining the impact of other peptide recognition domains on PPIN connectivity. I found that while protein BLAST scores, often used by network alignment methods, are effective at predicting protein orthology, they are not effective at predicting PPI conservation. I also find that the rate of interaction conservation is quite low, and that duplicated genes, which

can retain high levels of sequence similarity, have even lower rates. Looking at SH3-mediated PPI data, I could not identify a pattern of interaction conservation, finding that putative conserved SH3 domains infrequently mediated conserved interactions.

Altogether, these results indicate to me that PPIN evolution is highly complex, with multiple mechanisms at work, and that to devise methods that effectively utilize interactomic data from multiple species, better understanding of these mechanisms is required. Chapter 4 represents my first effort to pivot in that direction. The Pairwise Protein Alignment Analysis Tool (PPAAT) is an HTML-based visualization tool for closely examining pairs of proteins and their network neighbourhoods. In the current state of the science, critical, in-depth analysis of network alignments is difficult due to the lack of usable tools and the large amount of data. By integrating network data, domain data, and orthology data, PPAAT offers researchers a tool to assess whether two proteins should be aligned or not, among other possible uses. As an initial entry as a data integration/visualization tool designed specifically for close-zoom, cross-species network analysis, PPAAT will hopefully promote new work in viewing, characterizing, and clarifying the observable artefacts of PPIN evolution, and assist us in understanding network evolution in greater depth.

5.2 Future Directions

5.2.1 Full Development and Deployment of PPAAT

In its current state, PPAAT is an immature software tool, more suitable for internal than public use. In addition to the additional scientific features detailed in 4.3.1 Additional Features to be Implemented, PPAAT requires significant development and testing before it can be deployed as a public web application. While designed to use HTML and JavaScript, PPAAT currently lacks an online infrastructure. Furthermore, there are design decisions to be made about PPAAT's future direction, whether it should be further developed as a visualization tool or an integrated analysis tool, or both. With the goal of PPAAT being to inspire and facilitate the usage of interactomic data in biological research, user research must be performed to understand what features would be useful to potential PPAAT users, so that PPAAT can be designed to fit into their workflow.

One major improvement needed by PPAAT that has other external applications is the development of a more informative scoring function for indicating the interologous coherence between two proteins. Currently, PPAAT shows simple counting statistics, but provides no context to determine the significance of the presented statistics. Absent this information, users are reliant upon their personal experience and intuition to determine whether a given PPAAT view is matched or unmatched enough to draw conclusions from.

One possible solution would be to develop an orthology confidence score using the observed interaction conservation rates between orthologous and non-orthologous protein pairs, perhaps generated by a Bayesian predictor. With PPAAT, this would provide users contextual information on the significance of the network data visualized before them. A scoring function of this sort may have utility beyond PPAAT too, however. Such a score could also be used directly in network alignment research, to evaluate the predictive power of a given alignment by quantifying the number of protein pairs whose neighbourhoods have been significantly aligned or even as a measure for network alignment methods to use explicitly, either in evaluation or during the alignment process.

5.2.2 Interaction/Edge Attributes

A curious asymmetry in network alignment is that nodes (proteins) have many attributes that serve in constructing and evaluating alignments, but edges (interactions) have none. This can incentivize network alignment methods to either prioritize correct node alignment and relegate edge alignment to an afterthought, as there are no "incorrect" edge alignments, or compromise node alignment quality in an overzealous effort to maximize the number of edges aligned. Given the empirically observed trade-off between node alignment and edge alignment, as described in 1.2.4.2 Network Topology Assessment, it is imperative for further network alignment development that the relative importance of node alignment and edge alignment be understood.

One way to resolve this asymmetry may be to catalogue and annotate PPI attributes. Doing so, creates the notion of "incorrect" edge alignments, if the annotated attributes of two aligned edges are mismatched or incompatible. This will at least partially fix the asymmetry between nodes and edges, so the trade-off between their alignment can be better assessed, and provide additional depth to the concept of interaction conservation. Additional possibilities for edge attribute sources include mediation by specific domain types, structural data, and known but uncatalogued

characterizations from well-studied pathways and complexes. While some of this data may be sparse, exploring the depths of the available data and developing working examples would be an important step towards thinking about PPINs in their entirety, rather than from the traditionally gene-based perspective.

5.2.3 Critical Assessment and Validation of Network Alignment Methods

While there is extensive literature describing network alignment methods, their uptake in the broader scientific community has been rather limited. I hypothesize that this is due to the "big data" nature of PPINs and network alignment, which, when combined with data quality concerns and unimpressive prediction performance, makes the use of network alignment in smaller scale projects currently difficult and of dubious benefit to researchers. However, there is apparently little interest in bridging this gap in the network alignment community; one prominent network alignment researcher once remarked to me that it was not their job to determine how biological insights could be derived from network alignments, and that was "the biologists' job."

I disagree. I believe that it is of primary importance that network alignment research begets useful biological applications, and furthermore that network alignment researchers must work to ensure that occurs. As a starting point, state-of-the-art network alignment methods should be assessed for their ability to make useful biological predictions, using other existing prediction methods as a comparative baseline. Comparing network alignment methods against other prediction methods, instead of each other,[140] will determine the scientific value *added* by network alignment to the current repertoire of bioinformatic tools. This assessment should be performed with detailed breakdowns, rather than just summary statistics, so different areas of strength and weakness can also be identified, such as, for example, if there are certain classes of proteins that network alignment is more or less effective on.

If this is demonstrably the case, that network alignment is more effective with certain parts of the proteome or interactome, then network alignment can be recalibrated to specifically focus on these. Such focus may be desirable regardless; a PPIN can be thought of as multiple interaction networks overlaid together, and by splitting the PPIN into these constituent subnetworks, we might eliminate noisy cross-talk between these networks and clear up evolutionary signals for alignment. For example, we could separate out gene regulatory networks for alignment, as regulatory interactions have up- and down-regulatory attributes that could guide alignment, or

core metabolic networks that we expect to be highly conserved between species, or different IINs if we expect that there are different evolutionary processes at work on these biophysically distinct interaction types. By first studying these systems in isolation, we can confidently determine the evolutionary mechanisms influencing each, and whether these mechanisms are generalizable or specific, which can then guide the development of network alignment methods designed to reveal these mechanisms at work.

Such a considered approach should not be seen as retreating from a difficult problem. By first simplifying the network alignment problem and refining alignment methods within such smaller, more tractable problem domains, we might ultimately end up in a better situation to understand and solve global network alignment across the entire interactome. Without critical reassessment and recalibration, network alignment research may become increasingly impractical, contributing little to our overall understanding of biological systems and evolution.

5.2.4 Modelling the Mechanisms of Network Evolution

Unlike with sequence alignment, which benefits from rich knowledge about how DNA and protein sequence evolution works,[206-209] in network alignment there has been little work devoted to developing models for network evolution. While several early methods attempted to elucidate evolutionary models, more recent work does not. Currently, network alignment research handles network evolution with very high-level approaches, perhaps targeting one or two vaguely understood network evolution phenomena. For example, HubAlign[76] utilizes nodes identified as important by their local network topologies and ModuleAlign[79] utilizes the modular organization of the proteome. There are no commonly cited network alignment analogues to the Jukes-Cantor[210] or Kimura[211] models of DNA evolution, nor the BLOSUM[212] and PAM[213] sequence alignment scoring matrices. Within social network theory, there are established mechanisms like triadic closure, local bridges, and assortative mixing, that can help explain the "evolution" of social networks over time.[214] In contrast, it is very difficult to evaluate whether network alignments are capturing biologically relevant phenomena, and which phenomena, or scientific artefacts.

Instead, it may be more productive to understand the basic mechanisms of network evolution and build up a comprehensive model thereof, then develop a network alignment method to express the information captured within. Simpler network alignment methods that concretely capture

known evolutionary phenomena could be iteratively improved to include more phenomena as they are found, and would provide certainty to users about what biological information is contained within their alignment results. Such a considered, iterative approach would also increase the synergy between network alignment and network evolution research: unexpected results in a network alignment could point towards new evolutionary phenomena to be investigated, and as the dynamics of these phenomena are explained they can be incorporated back into new network alignment methods.

To begin this process requires comprehensive consideration of the accumulated network evolution information that currently exists. While some of this information resides within explicit network evolution research, much of it may also reside in protein complex research, protein structure research, and research focused on known modules and key proteins of interest. Much of this research contains observations relevant to the dynamics of network evolution, such as how a specific protein complex has evolved between species,[215] but this information has not been viewed through the perspective of network alignment. Though it may be a time-consuming process that may require substantive collaboration with other scientists working in fields other than network alignment, theorizing the phenomenological patterns that would emerge from known evolutionary dynamics or reconstructing the patterns that have already been observed could root network alignment more firmly within established fields of knowledge, providing traction for more innovation and applications in future network alignment research.

References

1. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53 (1970).

2. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7 (1981).

3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

4. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).

5. Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol* **19**, 341-8 (2009).

6. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

7. Pitre, S. *et al.* PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics* **7**, 365 (2006).

8. Ding, Z. & Kihara, D. Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features. *Curr Protoc Protein Sci* **93**, e62 (2018).

9. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).

10. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).

11. Gilson, M.K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* **44**, D1045-53 (2016).

12. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91 (2000).

13. Goel, R., Harsha, H.C., Pandey, A. & Prasad, T.S. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol Biosyst* **8**, 453-63 (2012).

14. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, D452-5 (2004).

15. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572-4 (2007).

16. Guldener, U. *et al.* MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* **34**, D436-41 (2006).

17. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-4 (2005).

18. Brown, K.R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076-82 (2005).

19. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. **45**, D362-d368 (2017).

20. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-d379 (2017).

21. Kim, J., Kim, I., Han, S.K., Bowie, J.U. & Kim, S. Network rewiring is an important mechanism of gene essentiality change. *Scientific Reports* **2**, 900 (2012).

22. Sharan, R. & Ideker, T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* **24**, 427-33 (2006).

23. Ryan, D.P. & Matthews, J.M. Protein-protein interactions in human disease. *Curr Opin Struct Biol* **15**, 441-6 (2005).

24. Gonzalez, M.W. & Kann, M.G. Chapter 4: Protein interactions and disease. *PLoS Comput Biol* **8**, e1002819 (2012).

25. Lage, K. Protein-protein interactions and genetic diseases: The interactome. *Biochim Biophys Acta* **1842**, 1971-1980 (2014).

26. Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* **9**, 1198-203 (1999).

27. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).

28. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **9 Suppl 1**, S4 (2008).

29. Wodak, S.J., Vlasblom, J., Turinsky, A.L. & Pu, S. Protein-protein interaction networks: the puzzling riches. *Curr Opin Struct Biol* **23**, 941-53 (2013).

30. Sambourg, L. & Thierry-Mieg, N. New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics* **11**, 605 (2010).

31. Edwards, A.M. *et al.* Too many roads not taken. *Nature* **470**, 163-5 (2011).

32. Edwards, A.M. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* **18**, 529-36 (2002).

33. Zhang, T. *et al.* Fancd2 in vivo interaction network reveals a non-canonical role in mitochondrial function. *Sci Rep* **7**, 45626 (2017).

34. Boucher, L. *et al.* (2017). *The Biological General Repository for Interaction Datasets* (Version 3.4.147) [Data set]. Retrieved from: https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.4.147/

35. McFarland, M.A., Ellis, C.E., Markey, S.P. & Nussbaum, R.L. Proteomics analysis identifies phosphorylation-dependent alpha-synuclein protein interactions. *Mol Cell Proteomics* **7**, 2123-37 (2008).

36. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40 (2008).

37. Piazzi, M. *et al.* Phosphoinositide-specific phospholipase C beta 1b (PI-PLCbeta1b) interactome: affinity purification-mass spectrometry analysis of PI-PLCbeta1b with nuclear protein. *Mol Cell Proteomics* **12**, 2220-35 (2013).

38. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res* **41**, D816-23 (2013).

39. Daley, D.O. The assembly of membrane proteins into complexes. *Curr Opin Struct Biol* **18**, 420-4 (2008).

40. Hunke, S. & Müller, V.S. Approaches to Analyze Protein-Protein Interactions of Membrane Proteins in *Protein Interactions* (eds. Cai, J. & Wang, R.E.) 327-348 (IntechOpen, 2012).

41. Petschnigg, J. *et al.* The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nat Methods* **11**, 585-92 (2014).

42. Smialowski, P. *et al.* The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res* **38**, D540-4 (2010).

43. Ali, W. & Deane, C.M. Evolutionary analysis reveals low coverage as the major challenge for protein interaction network alignment. *Mol Biosyst* **6**, 2296-304 (2010).

44. Johnson, M.E. & Hummer, G. Interface-Resolved Network of Protein-Protein Interactions. *PLoS Comput Biol* **9**, e1003065 (2013).

45. Han, J.D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93 (2004).

46. Xin, X. *et al.* SH3 interactome conserves general function over specific form. *Mol Syst Biol* **9**, 652 (2013).

47. Reimand, J., Hui, S., Jain, S., Law, B. & Bader, G.D. Domain-mediated protein interaction prediction: From genome to network. *FEBS Lett* **586**, 2751-63 (2012).

48. Reimand, J., Wagih, O. & Bader, G.D. The mutational landscape of phosphorylation signaling in cancer. *Sci Rep* **3**, 2651 (2013).

49. Tonikian, R. *et al.* Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* **7**, e1000218 (2009).

50. Teyra, J. *et al.* Comprehensive Analysis of the Human SH3 Domain Family Reveals a Wide Variety of Non-canonical Specificities. *Structure* (2017).

51. Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **21**, 3401-8 (2005).

52. Wernicke, S. & Rasche, F. Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics* **23**, 1978-85 (2007).

53. Cheng, Q., Harrison, R. & Zelikovsky, A. MetNetAligner: a web service tool for metabolic network alignments. *Bioinformatics* **25**, 1989-90 (2009).

54. Ficklin, S.P. & Feltus, F.A. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiol* **156**, 1244-56 (2011).

55. Singh, R., Xu, J. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci U S A* **105**, 12763-8 (2008).

56. Babai, L. Graph isomorphism in quasipolynomial time. in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* 684-697 (ACM, 2016).

57. Babai, L. Fixing the UPCC case of Split-or-Johnson. *Graph Isomorphism update, January 9, 2017.* Available at http://people.cs.uchicago.edu/~laci/upcc-fix.pdf (2017).

58. Cook, S.A. The Complexity of Theorem Proving Procedures. *Proc Third Annual ACM Symposium on Theory of Computing*, 151-158 (1971).

59. Emmert-Streib, F., Dehmer, M. & Shi, Y. Fifty years of graph matching, network alignment and network comparison. *Information Sciences* **346-347**, 180-197 (2016).

60. Kelley, B.P. *et al.* PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res* **32**, W83-8 (2004).

61. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85 (2004).

62. Kalaev, M., Smoot, M., Ideker, T. & Sharan, R. NetworkBLAST: comparative analysis of protein networks. *Bioinformatics* **24**, 594-6 (2008).

63. Pache, R.A. & Aloy, P. A novel framework for the comparative analysis of biological networks. *PLoS One* **7**, e31220 (2012).

64. Koyuturk, M. *et al.* Pairwise alignment of protein interaction networks. *J Comput Biol* **13**, 182-99 (2006).

65. Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H. & Batzoglou, S. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res* **16**, 1169-81 (2006).

66. Ciriello, G., Mina, M., Guzzi, P.H., Cannataro, M. & Guerra, C. AlignNemo: a local network alignment method to integrate homology and topology. *PLoS One* **7**, e38107 (2012).

67. Mina, M. & Guzzi, P.H. Improving the Robustness of Local Network Alignment: Design and Extensive Assessment of a Markov Clustering-Based Approach. *IEEE/ACM Trans Comput Biol Bioinform* **11**, 561-72 (2014).

68. Micale, G., Pulvirenti, A., Giugno, R. & Ferro, A. GASOLINE: a Greedy And Stochastic algorithm for optimal Local multiple alignment of Interaction NEtworks. *PLoS One* **9**, e98750 (2014).

69. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. (Stanford InfoLab, 1999).

70. Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W. & Przulj, N. Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface* **7**, 1341-54 (2010).

71. Memisevic, V. & Przulj, N. C-GRAAL: common-neighbors-based global GRAph ALignment of biological networks. *Integr Biol (Camb)* **4**, 734-43 (2012).

72. Kuchaiev, O. & Przulj, N. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* **27**, 1390-6 (2011).

73. Patro, R. & Kingsford, C. Global network alignment using multiscale spectral signatures. *Bioinformatics* **28**, 3105-14 (2012).

74. Meng, L., Crawford, J., Striegel, A. & Milenkovic, T. IGLOO: Integrating global and local biological network alignment. Available at https://ui.adsabs.harvard.edu/#abs/2016arXiv160406111M (2016).

75. Neyshabur, B., Khadem, A., Hashemifar, S. & Arab, S.S. NETAL: a new graph-based method for global alignment of protein-protein interaction networks. *Bioinformatics* **29**, 1654-62 (2013).

76. Hashemifar, S. & Xu, J. HubAlign: an accurate and efficient method for global alignment of protein-protein interaction networks. *Bioinformatics* **30**, i438-44 (2014).

77. Milenkovic, T., Ng, W.L., Hayes, W. & Przulj, N. Optimal network alignment with graphlet degree vectors. *Cancer Inform* **9**, 121-37 (2010).

78. Chindelevitch, L., Ma, C.-Y., Liao, C.-S. & Berger, B. Optimizing a global alignment of protein interaction networks. *Bioinformatics* **29**, 2765-2773 (2013).

79. Hashemifar, S., Ma, J., Naveed, H., Canzar, S. & Xu, J. ModuleAlign: module-based global alignment of protein-protein interaction networks. *Bioinformatics* **32**, i658-i664 (2016).

80. Mamano, N. & Hayes, W.B. SANA: simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics* **33**, 2156-2164 (2017).

81. Clark, C. & Kalita, J. A multiobjective memetic algorithm for PPI network alignment. *Bioinformatics* **31**, 1988-98 (2015).

82. Ibragimov, R., Malek, M., Guo, J. & Baumbach, J. *GEDEVO: An evolutionary graph edit distance algorithm for biological network alignment*, (2013).

83. Vijayan, V., Saraph, V. & Milenkovic, T. MAGNA++: Maximizing Accuracy in Global Network Alignment via both node and edge conservation. *Bioinformatics* **31**, 2409-11 (2015).

84. Ibragimov, R., Malek, M., Baumbach, J. & Guo, J. Multiple graph edit distance: simultaneous topological alignment of multiple protein-protein interaction networks with an evolutionary algorithm. in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* 277-284 (ACM, Vancouver, BC, Canada, 2014).

85. Vijayan, V. & Milenkovic, T. Multiple network alignment via multiMAGNA++. Available at https://ui.adsabs.harvard.edu/#abs/2016arXiv160401740V (2016).

86. Liao, C.S., Lu, K., Baym, M., Singh, R. & Berger, B. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25**, i253-8 (2009).

87. Malod-Dognin, N. & Przulj, N. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics* **31**, 2182-9 (2015).

88. Zaslavskiy, M., Bach, F. & Vert, J.P. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics* **25**, i259-67 (2009).

89. El-Kebir, M., Heringa, J. & Klau, G. Natalie 2.0: Sparse Global Network Alignment as a Special Case of Quadratic Assignment. *Algorithms* **8**, 1035 (2015).

90. Gligorijevic, V., Malod-Dognin, N. & Przulj, N. Fuse: multiple network alignment via data fusion. *Bioinformatics* **32**, 1195-203 (2016).

91. Bandyopadhyay, S., Sharan, R. & Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res* **16**, 428-35 (2006).

92. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

93. The Gene Ontology Consortium. *How can I calculate the "level" of a GO term? | Gene Ontology Consortium*. (2018) Available at http://www.geneontology.org/faq/how-can-i-calculate-level-go-term (Accessed: 02-10-2018)

94. Alterovitz, G., Xiang, M., Mohan, M. & Ramoni, M.F. GO PaD: the Gene Ontology Partition Database. *Nucleic Acids Res* **35**, D322-7 (2007).

95. Carbon, S. & Mungall, C. (2018). *Gene Ontology Data Archive* (Version 2018-12-01) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1899458.

96. Lord, P.W., Stevens, R.D., Brass, A. & Goble, C.A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275-83 (2003).

97. Pesquita, C. Semantic Similarity in the Gene Ontology. in *The Gene Ontology Handbook* (eds. Dessimoz, C. & Škunca, N.) 161-173 (Springer New York, New York, NY, 2017).

98. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

99. Saraph, V. & Milenkovic, T. MAGNA: Maximizing Accuracy in Global Network Alignment. *Bioinformatics* **30**, 2931-40 (2014).

100. Malod-Dognin, N., Ban, K. & Przulj, N. Unified Alignment of Protein-Protein Interaction Networks. *Sci Rep* **7**, 953 (2017).

101. Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-50 (2007).

102. Beltrao, P. & Serrano, L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol* **3**, e25 (2007).

103. Shou, C. *et al.* Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol* **7**, e1001050 (2011).

104. Lewis, A.C., Jones, N.S., Porter, M.A. & Deane, C.M. What evidence is there for the homology of protein-protein interactions? *PLoS Comput Biol* **8**, e1002645 (2012).

105. Sun, M.G., Sikora, M., Costanzo, M., Boone, C. & Kim, P.M. Network evolution: rewiring and signatures of conservation in signaling. *PLoS Comput Biol* **8**, e1002411 (2012).

106. Aladag, A.E. & Erten, C. SPINAL: scalable protein interaction network alignment. *Bioinformatics* **29**, 917-24 (2013).

107. Sun, M.G. & Kim, P.M. Evolution of biological interaction networks: from models to real data. *Genome Biol* **12**, 235 (2011).

108. Kim, P.M., Korbel, J.O. & Gerstein, M.B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-9 (2007).

109. Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938-41 (2006).

110. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-2 (2001).

111. Kim, P.M., Sboner, A., Xia, Y. & Gerstein, M. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol* **4**, 179 (2008).

112. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59 (2007).

113. Janjic, V., Sharan, R. & Przulj, N. Modelling the yeast interactome. *Sci Rep* **4**, 4273 (2014).

114. Walhout, A.J. *et al.* Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287**, 116-22 (2000).

115. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **11**, 2120-6 (2001).

116. Folador, E.L. *et al.* An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. *Integr Biol (Camb)* **6**, 1080-7 (2014).

117. Brown, K.R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol* **8**, R95 (2007).

118. Nguyen, P.V., Srihari, S. & Leong, H.W. Identifying conserved protein complexes between species by constructing interolog networks. *BMC Bioinformatics* **14 Suppl 16**, S8 (2013).

119. Kappei, D. *et al.* Phylointeractomics reconstructs functional evolution of protein binding. *Nat Commun* **8**, 14334 (2017).

120. Agarwal, S., Deane, C.M., Porter, M.A. & Jones, N.S. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol* **6**, e1000817 (2010).

121. Gillis, J. & Pavlidis, P. The impact of multifunctional genes on "guilt by association" analysis. *PLoS One* **6**, e17258 (2011).

122. Gillis, J. & Pavlidis, P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. *PLoS Computational Biology* **8**, e1002444 (2012).

123. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-13 (2004).

124. Przulj, N., Corneil, D.G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508-15 (2004).

125. Ispolatov, I., Krapivsky, P.L. & Yuryev, A. Duplication-divergence model of protein interaction network. *Physical review. E, Statistical, nonlinear, and soft matter physics* **71**, 061911-061911 (2005).

126. Taylor, J.S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**, 615-43 (2004).

127. Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S. & Koonin, E.V. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol* **2**, 18 (2002).

128. Thorne, T. & Stumpf, M.P. Graph spectral analysis of protein interaction network evolution. *J R Soc Interface* **9**, 2653-66 (2012).

129. Wagner, A. How the global structure of protein interaction networks evolves. *Proc Biol Sci* **270**, 457-66 (2003).

130. Evlampiev, K. & Isambert, H. Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proceedings of the National Academy of Sciences* **105**, 9863-9868 (2008).

131. Ratmann, O., Wiuf, C. & Pinney, J.W. From evidence to inference: Probing the evolution of protein interaction networks. *HFSP Journal* **3**, 290-306 (2009).

132. Flannick, J., Novak, A., Do, C.B., Srinivasan, B.S. & Batzoglou, S. Automatic parameter learning for multiple local network alignment. *J Comput Biol* **16**, 1001-22 (2009).

133. Berg, J., Lässig, M. & Wagner, A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology* **4**, 51 (2004).

134. Batagelj, V. & Mrvar, A. Pajek-program for large network analysis. *Connections* **21**, 47-57 (1998).

135. Wiese, R., Eiglsperger, M. & Kaufmann, M. yFiles: Visualization and Automatic Layout of Graphs. 453-454 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2002).

136. Brown, K.R. *et al.* NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* **25**, 3327-9 (2009).

137. Hao, Y. *et al.* OrthoNets: simultaneous visual analysis of orthologs and their interaction neighborhoods across different organisms. *Bioinformatics* **27**, 883-4 (2011).

138. Micale, G., Continella, A., Ferro, A., Giugno, R. & Pulvirenti, A. GASOLINE: a Cytoscape app for multiple local alignment of PPI networks. *F1000Res* **3**, 140 (2014).

139. Ge, H., Walhout, A.J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19**, 551-60 (2003).

140. Clark, C. & Kalita, J. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* **30**, 2351-2359 (2014).

141. Tong, A.H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321-4 (2002).

142. Mulzer, W. & Rote, G. Minimum-weight triangulation is NP-hard. *Journal of the ACM* **55**, 1-29 (2008).

143. Keshava Prasad, T.S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-72 (2009).

144. Przulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177-83 (2007).

145. Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **21**, 3401-3408 (2005).

146. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-14 (2012).

147. El-Kebir, M., Heringa, J. & Klau, G.W. Lagrangian Relaxation Applied to Sparse Global Network Alignment. *arXiv:1108.4358* (2011).

148. Jain, S. & Bader, G.D. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* **11**, 562 (2010).

149. Soulard, A. *et al.* Saccharomyces cerevisiae Bzz1p is implicated with type I myosins in actin patch polarization and is able to recruit actin-polymerizing machinery in vitro. *Mol Cell Biol* **22**, 7889-906 (2002).

150. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. Interaction databases on the same page. *Nat Biotechnol* **29**, 391-3 (2011).

151. Giuliani, C. *et al.* Requirements for F-BAR Proteins TOCA-1 and TOCA-2 in Actin Dynamics and Membrane Trafficking during Caenorhabditis elegans Oocyte Growth and Embryonic Epidermal Morphogenesis. *PLoS Genetics* **5**, e1000675 (2009).

152. Bu, W. *et al.* Cdc42 Interaction with N-WASP and Toca-1 Regulates Membrane Tubulation, Vesicle Formation and Vesicle Motility: Implications for Endocytosis. *PLoS ONE* **5**, e12153 (2010).

153. Todor, A., Dobra, A. & Kahveci, T. Probabilistic biological network alignment. *IEEE/ACM Trans Comput Biol Bioinform* **10**, 109-21 (2013).

154. Berg, J. & Lassig, M. Cross-species analysis of biological networks by Bayesian alignment. *Proc Natl Acad Sci U S A* **103**, 10967-72 (2006).

155. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).

156. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84-90 (2012).

157. Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**, D196-203 (2010).

158. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89 (2003).

159. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54-61 (2007).

160. Moustafa, A. *JAligner: Open source Java implementation of Smith-Waterman*. (2012) Available at http://jaligner.sourceforge.net/ (Accessed: 26th May, 2015)

161. Assenov, Y., Ramirez, F., Schelhorn, S.E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282-4 (2008).

162. Wang, P.I. & Marcotte, E.M. It's the machine that matters: Predicting gene function and phenotype from protein networks. *J Proteomics* **73**, 2277-89 (2010).

163. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).

164. Lo, Y.S., Huang, S.H., Luo, Y.C., Lin, C.Y. & Yang, J.M. Reconstructing genome-wide protein-protein interaction networks using multiple strategies with homologous mapping. *PLoS One* **10**, e0116347 (2015).

165. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-61 (2003).

166. Mika, S. & Rost, B. Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol* **2**, e79 (2006).

167. Wagner, A. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol* **19**, 1760-8 (2002).

168. Darby, C.A., Stolzer, M., Ropp, P.J., Barker, D. & Durand, D. Xenolog classification. *Bioinformatics* **33**, 640-649 (2017).

169. Sahni, N. *et al.* Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* **23**, 649-57 (2013).

170. Song, N., Joseph, J.M., Davis, G.B. & Durand, D. Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins. *PLOS Computational Biology* **4**, e1000063 (2008).

171. Vogel, C., Teichmann, S.A. & Pereira-Leal, J. The relationship between domain duplication and recombination. *J Mol Biol* **346**, 355-65 (2005).

172. Chothia, C., Gough, J., Vogel, C. & Teichmann, S.A. Evolution of the protein repertoire. *Science* **300**, 1701-3 (2003).

173. Vogel, C. & Chothia, C. Protein family expansions and biological complexity. *PLoS Comput Biol* **2**, e48 (2006).

174. Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445-52 (2003).

175. Jin, J. *et al.* Eukaryotic protein domains as functional units of cellular evolution. *Sci Signal* **2**, ra76 (2009).

176. Ghoshal, G., Chi, L. & Barabasi, A.L. Uncovering the role of elementary processes in network evolution. *Sci Rep* **3**, 2920 (2013).

177. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**, 1283-92 (2001).

178. Yu, H. *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* **14**, 1107-18 (2004).

179. Zarrinpar, A., Park, S.H. & Lim, W.A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676-80 (2003).

180. Tan, C.S. *et al.* Positive selection of tyrosine loss in metazoan evolution. *Science* **325**, 1686-8 (2009).

181. Moore, A.D. & Bornberg-Bauer, E. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol* **29**, 787-96 (2012).

182. Lim, W.A. & Pawson, T. Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* **142**, 661-7 (2010).

183. Law, B. & Bader, G.D. GreedyPlus: An Algorithm for the Alignment of Interface Interaction Networks. *Sci Rep* **5**, 12074 (2015).

184. Guipponi, M. *et al.* Two isoforms of a human intersectin (ITSN) protein are produced by brain-specific alternative splicing in a stop codon. *Genomics* **53**, 369-76 (1998).

185. Yamabhai, M. *et al.* Intersectin, a novel adaptor protein with two Eps15 homology and five Src homology 3 domains. *J Biol Chem* **273**, 31401-7 (1998).

186. Teckchandani, A., Mulkearns, E.E., Randolph, T.W., Toida, N. & Cooper, J.A. The clathrin adaptor Dab2 recruits EH domain scaffold proteins to regulate integrin beta1 endocytosis. *Mol Biol Cell* **23**, 2905-16 (2012).

187. Malacombe, M. *et al.* Intersectin-1L nucleotide exchange factor regulates secretory granule exocytosis by activating Cdc42. *The EMBO Journal* **25**, 3494-3503 (2006).

188. Jain, S. & Bader, G.D. Predicting physiologically relevant SH3 domain mediated protein-protein interactions in yeast. *Bioinformatics* **32**, 1865-72 (2016).

189. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-9 (2012).

190. Schoenrock, A. *et al.* Evolution of protein-protein interaction networks in yeast. *PLoS One* **12**, e0171920 (2017).

191. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**, D514-D517 (2005).

192. Finn, R.D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-85 (2016).

193. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710-6 (2016).

194. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W. & Bruford, E.A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* **43**, D1079-85 (2015).

195. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-12 (2015).

196. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **35**, D26-31 (2007).

197. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-45 (2016).

198. Sigrist, C.J. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res* **41**, D344-7 (2013).

199. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* **43**, D257-60 (2015).

200. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**, 903-19 (2001).

201. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**, D213-21 (2015).

202. Castagnoli, L. *et al.* Selectivity and promiscuity in the interaction network mediated by protein recognition modules. *FEBS Lett* **567**, 74-9 (2004).

203. Sonnhammer, E.L. & Ostlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43**, D234-9 (2015).

204. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**, D363-8 (2006).

205. Kim, T. *et al.* MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res* **40**, e47 (2012).

206. Kimura, M. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet* **66**, 367-86 (1991).

207. Delport, W., Scheffler, K. & Seoighe, C. Models of coding sequence evolution. *Brief Bioinform* **10**, 97-109 (2009).

208. Pal, C., Papp, B. & Lercher, M.J. An integrated view of protein evolution. *Nat Rev Genet* **7**, 337-48 (2006).

209. Koonin, E.V. Are there laws of genome evolution? *PLoS Comput Biol* **7**, e1002173 (2011).

210. Jukes, T.H. & Cantor, C.R. Evolution of protein molecules. *Mammalian protein metabolism* **3**, 132 (1969).

211. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-20 (1980).

212. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-9 (1992).

213. Dayhoff, M.O. Atlas of protein sequence and structure. (1965).

214. Easley, D. & Kleinberg, J. *Networks, crowds, and markets: Reasoning about a highly connected world*, (Cambridge University Press, 2010).

215. Marsh, J.A. & Teichmann, S.A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem* **84**, 551-75 (2015).

Copyright Acknowledgements

Appendices

.

A Additional figures demonstrating the impact of gene duplication on interolog conservation

In 3.2.1 - Interolog conservation across species, the impact of gene duplication on the conservation of PPIs is discussed. Figure 3-4 and Figure 3-5 showed that the rate of conservation of PPIs from *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus* in the *H. sapiens* interactome is greatly impacted by the duplication of genes. Figure A-1 and Figure A-2 show the complete pairwise comparisons between all five species.

While Table 3-2 showed the percentage of PPIs that had at least 1 interolog in another species, the mean rate of conservation for any given PPI is lower due to the fact that many PPIs are not fully conserved amongst all possible interologs. This is possibly due to the loss of one or more "copies" of the original interaction after duplication or the PPI is a novel gain in the origin species. Furthermore, because some genes are highly duplicated, these genes and their protein products, with their lower rates of interaction conservation, have an outsized impact on the overall rate of interaction conservation, creating a large source of possible error for scientific methods that operate on PPIs with the presumption of conservation.

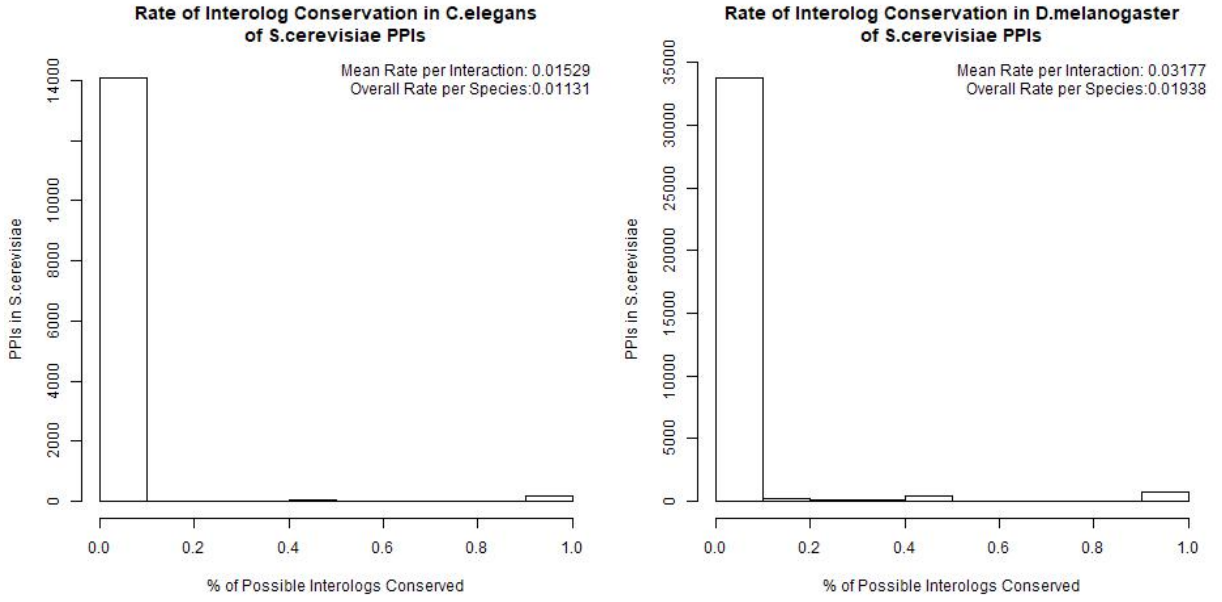


Figure A-1 - % interolog conservation for different species in various model interactomes.
(Continued on next page.)

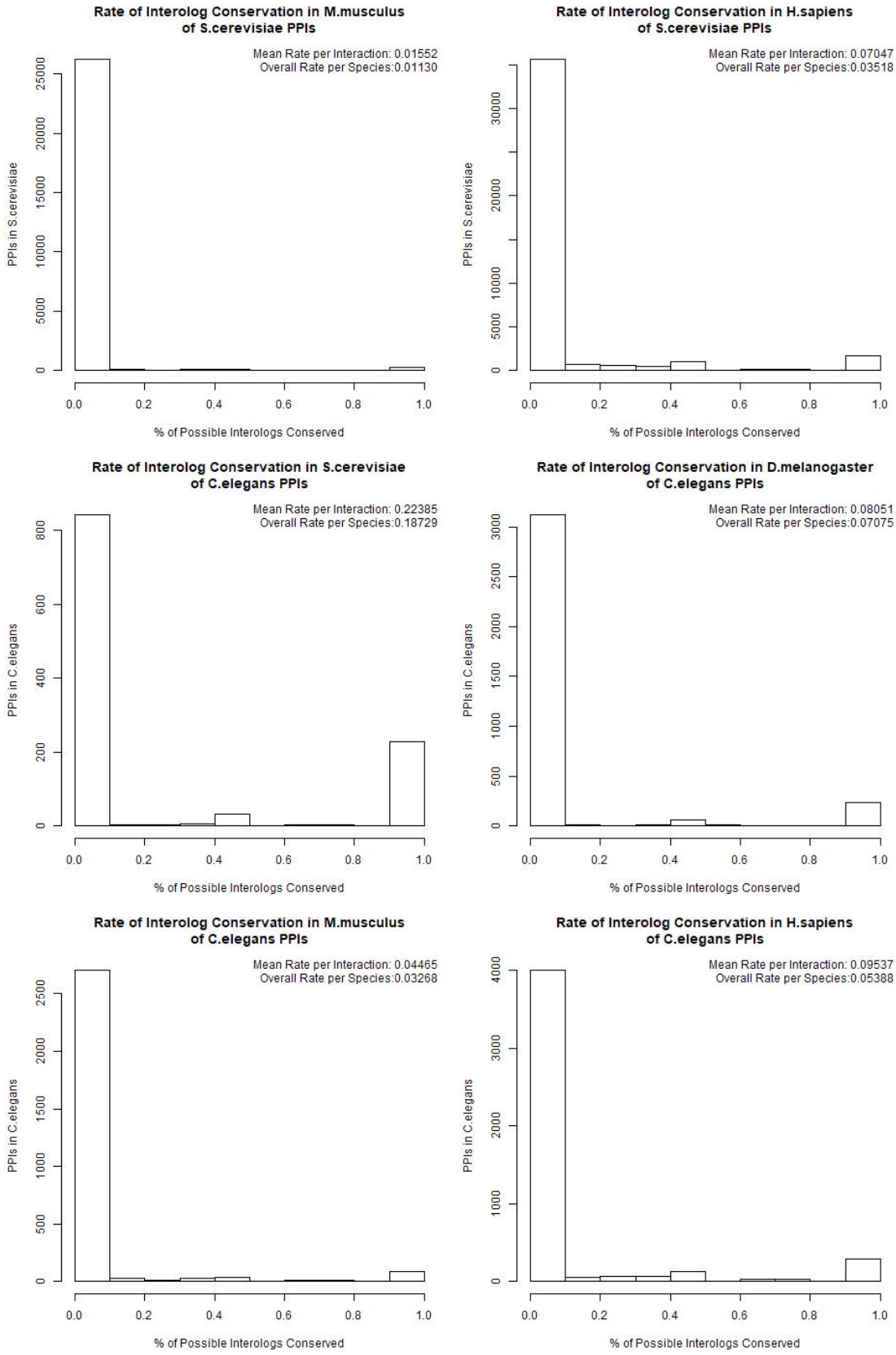


Figure A-1 - % interolog conservation for different species in various model interactomes.
(Continued on next page.)

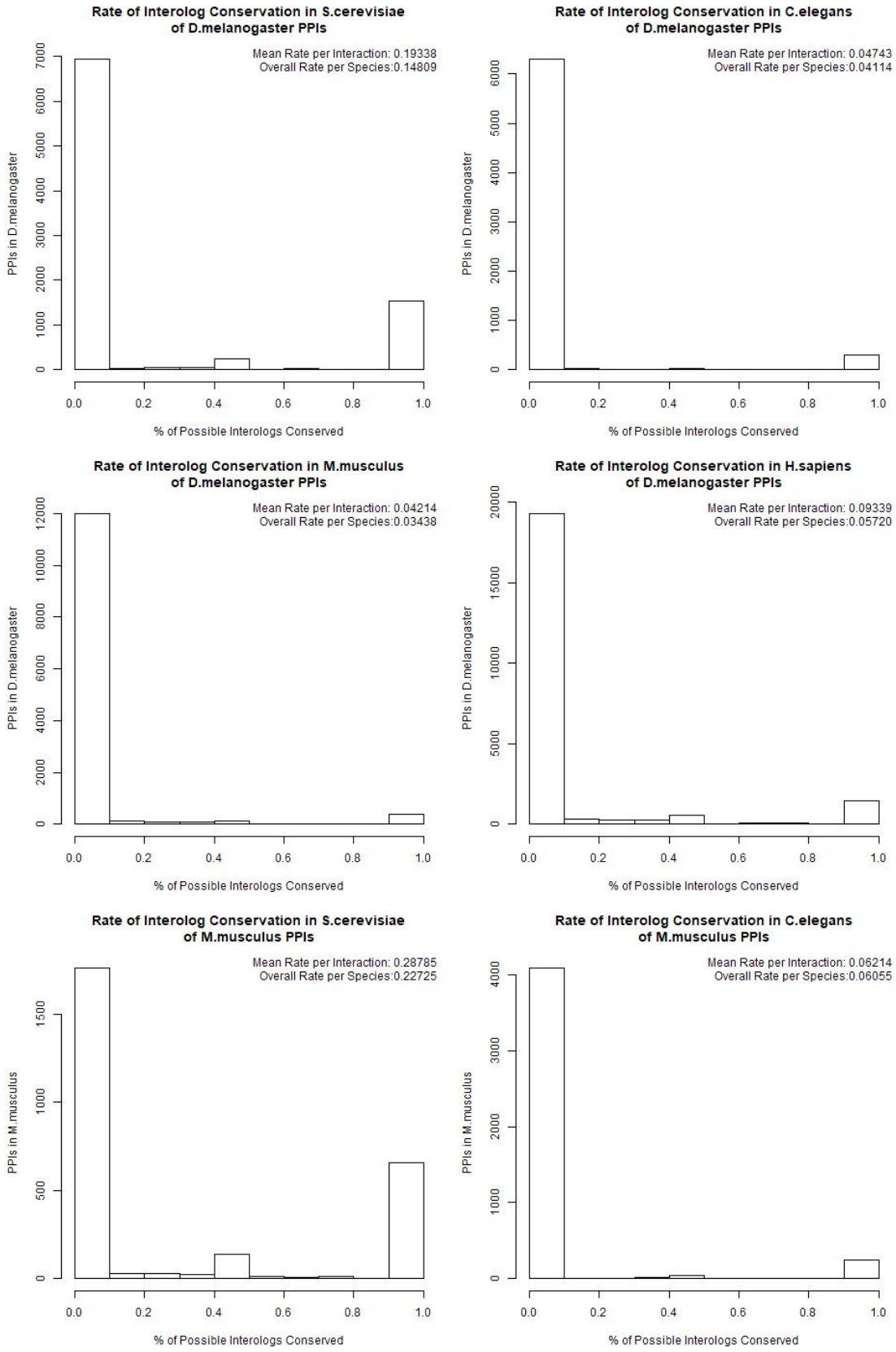


Figure A-1 - % interolog conservation for different species in various model interactomes.
(Continued on next page.)

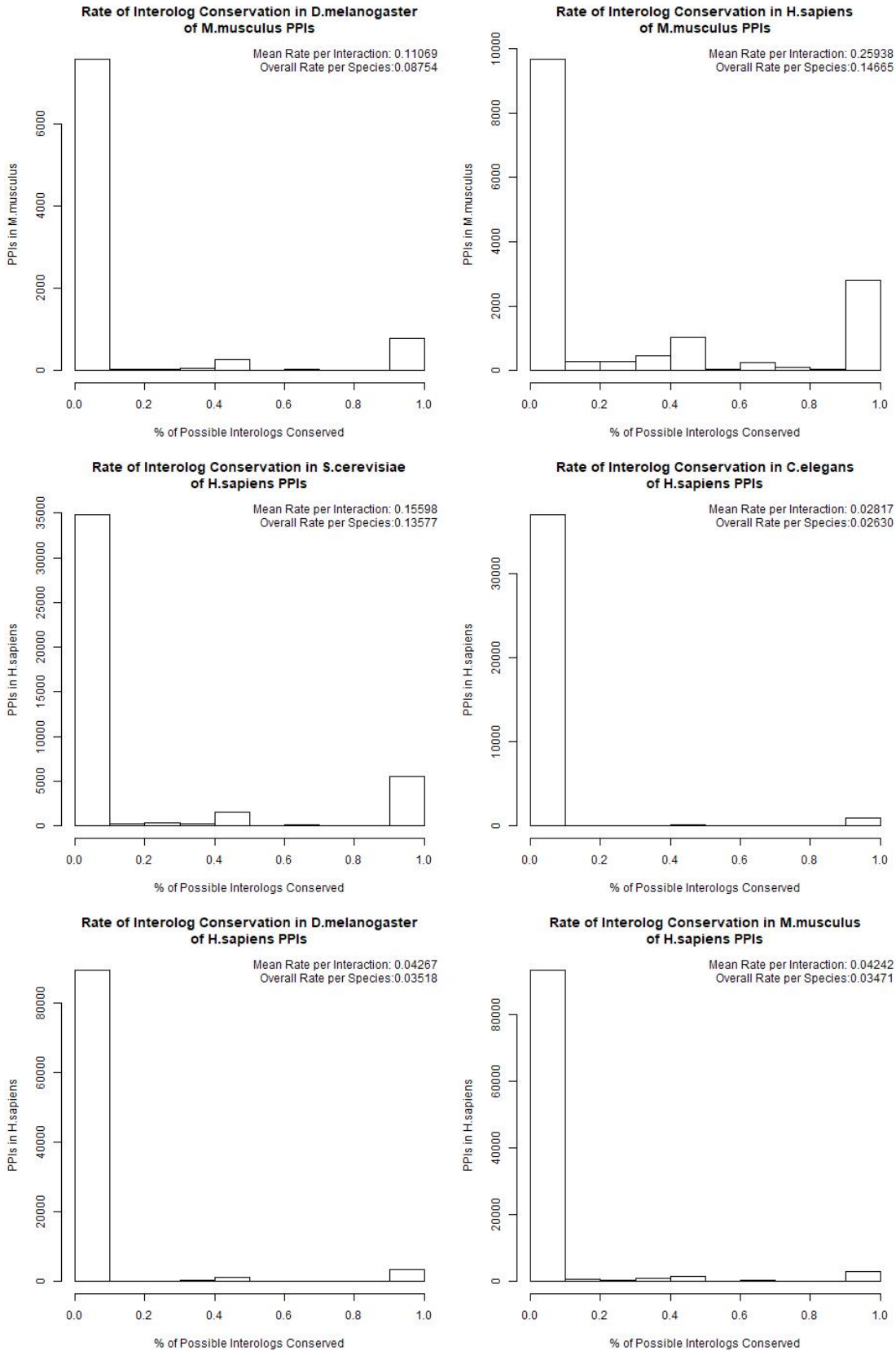


Figure A-1 - % interolog conservation for different species in various model interactomes.
(Caption on next page.)

Figure A-1 - % interolog conservation for different species in various model interactomes. This figure shows the distribution of conservation rates per PPI in the origin species, where a value of 1.0 indicates that the PPI is conserved between *all* orthologs of the protein interactors in a target species, and 0.0 indicates that no interaction is ever found between any of the target species orthologs. Compared to Table 3-2, which presented the % of PPI that had *any* conserved interolog in the human interactome, these rates of conservation are much lower, as there are often multiple orthologs in the target species for any given protein in the origin species. In the top right are listed the mean rate of conservation for a given PPI in the origin species amongst all possible interologs in the target interactome, and the total rate of conservation of all possible interologs in the target species.

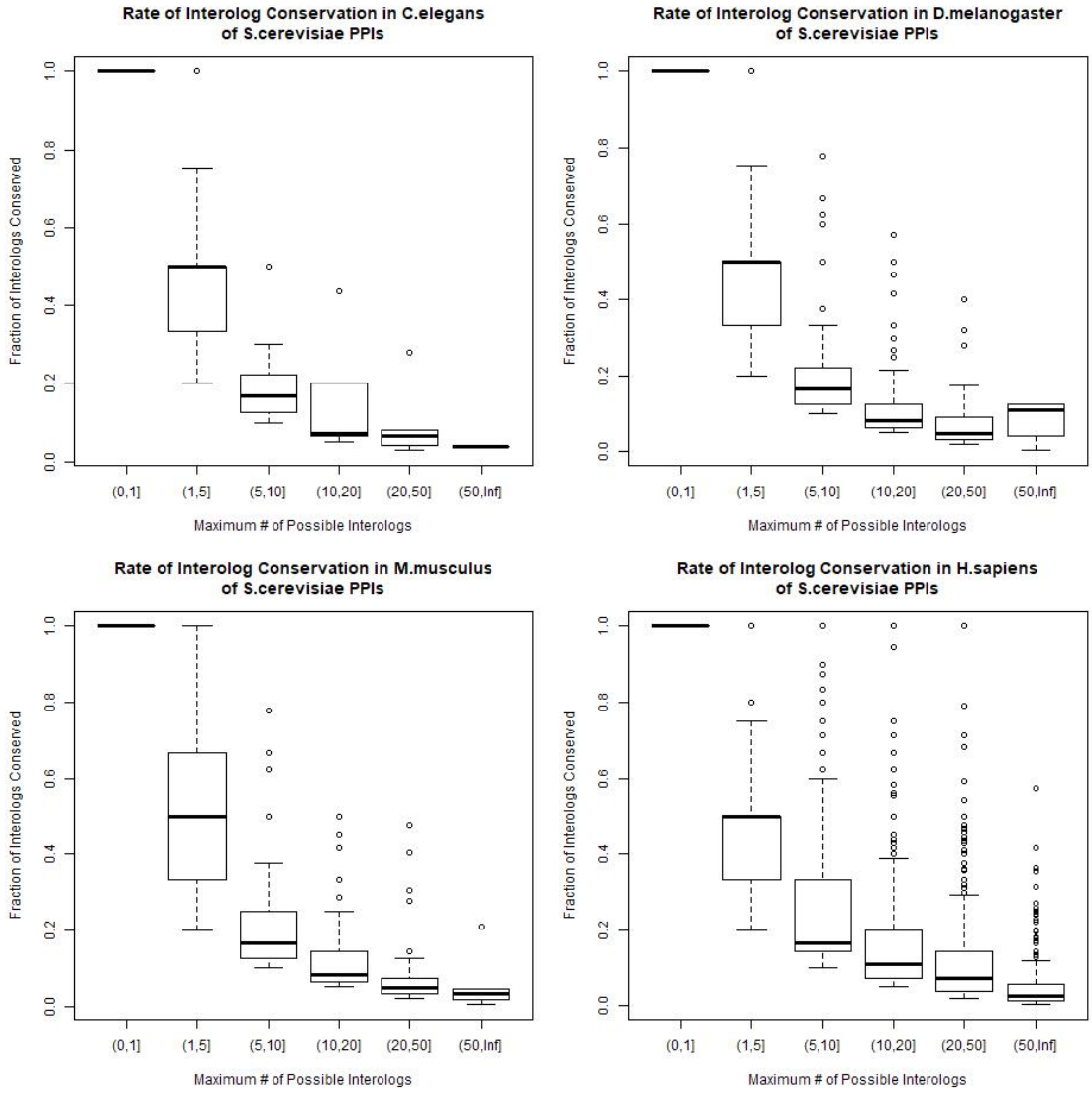


Figure A-2 – % interolog conservation for PPIs with known interologs, normalized by the number of potential interologs. (Continued on next page.)

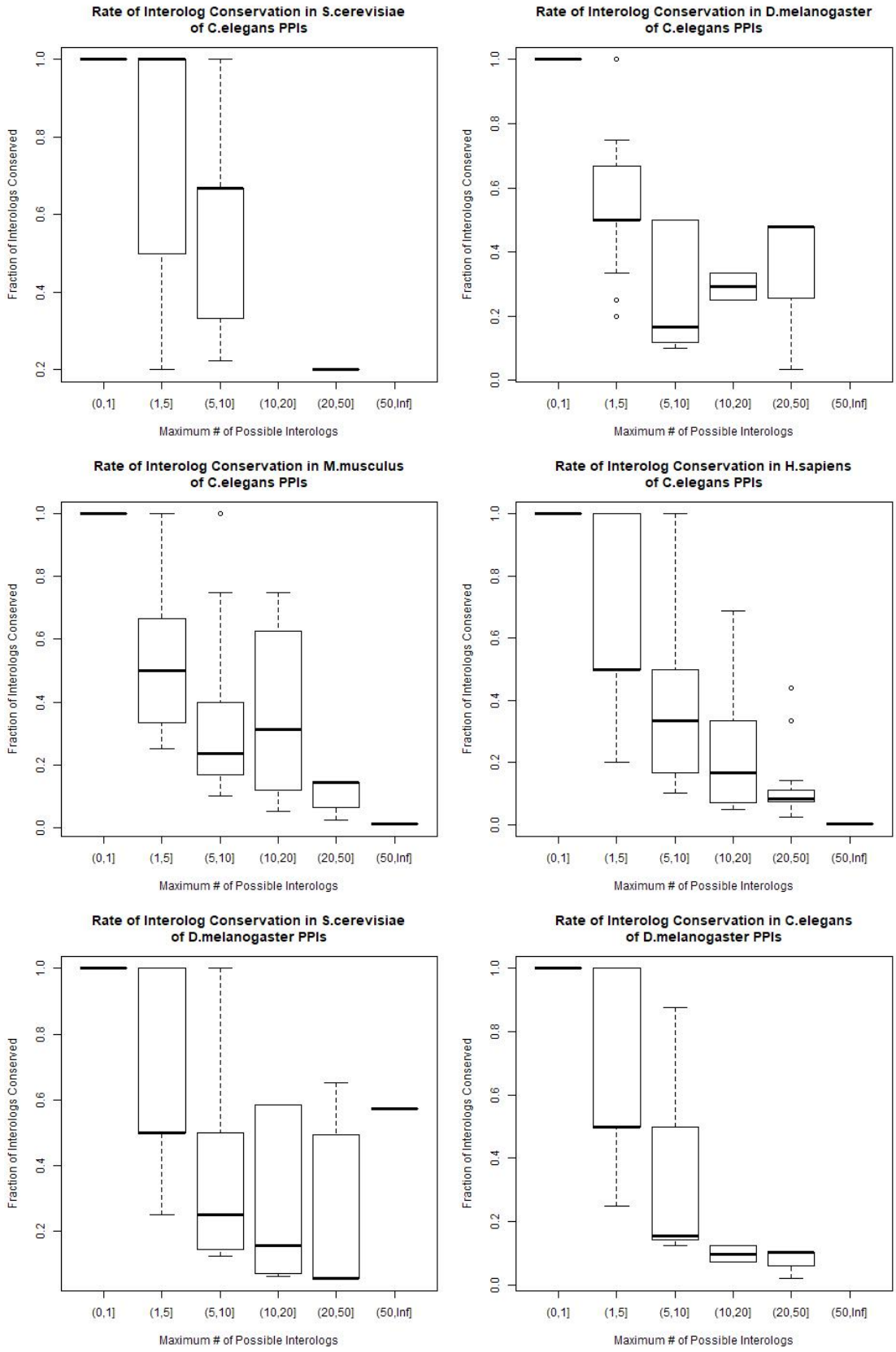


Figure A-2 – % interolog conservation for PPIs with known interologs, normalized by the number of potential interologs. (Continued on next page.)

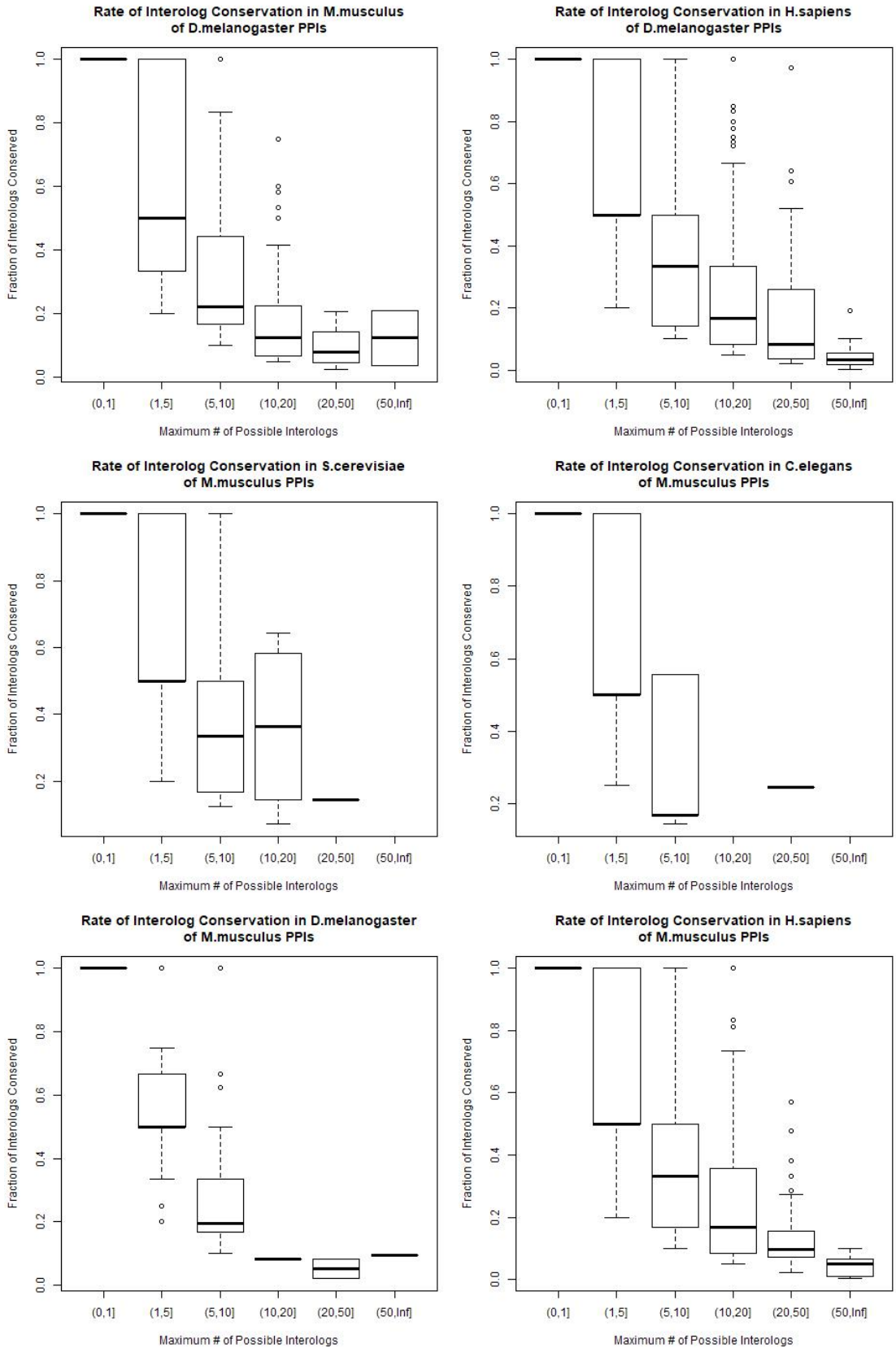


Figure A-2 – % interolog conservation for PPIs with known interologs, normalized by the number of potential interologs. (Continued on next page.)

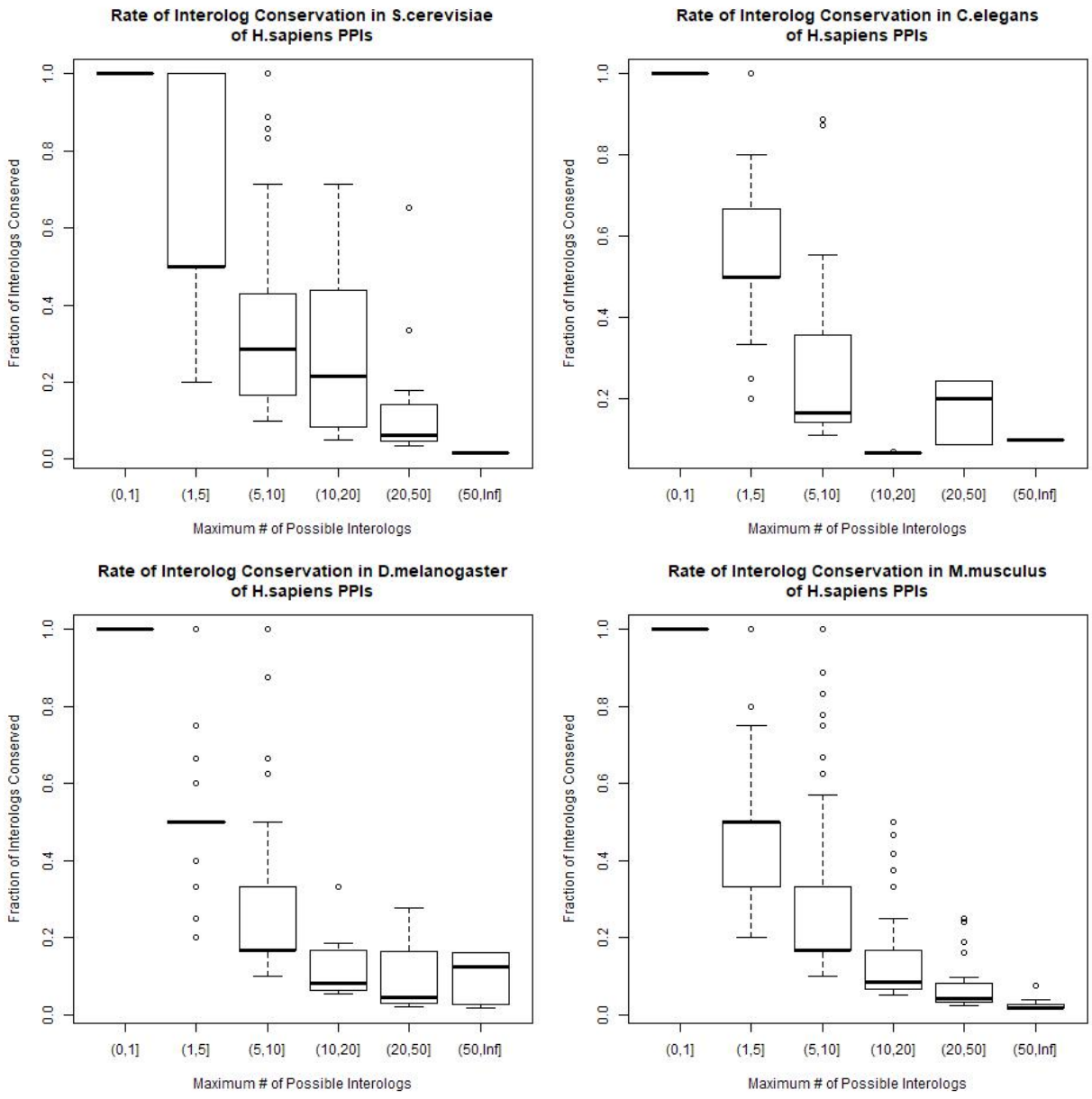


Figure A-2 – % interolog conservation for PPIs with known interologs, normalized by the number of potential interologs. Due to gene duplication events resulting in many-to-many ortholog relationships, some PPIs can be conserved more than once, resulting in multiple potential interologs. These histograms show the level of interolog conservation of PPIs, grouped by the maximum number of potential interologs that could have been found based on the number of orthologs in the target species. PPIs with no interologs were excluded, as they overwhelmed all other PPIs due to low overall rates of interolog conservation.

B Protein-protein interactions of human intersectin and its orthologs

In 3.2.3 - The importance of domain architecture in protein-protein interaction conservation, the known SH3 interaction data for *H. sapiens* (human) intersectin proteins (ITSN1 and ITSN2) and their orthologs in *C. elegans* (worm, ITSN-1) and *S. cerevisiae* (yeast, EDE1) were analyzed. The protein-protein interactors for these proteins are listed below, using PPI data from iRefIndex, with SH3-specific interactions split out. The worm[46] and yeast[49] SH3 binding data were experimentally derived using phage display and their targets subsequently verified with yeast-two-hybrid. The human SH3 binding data[50] was predicted based on phage display experiments.

Table B-1 – The protein interactors for human ITSN1 and ITSN2, as well as their yeast and worm orthologs.

Protein/Domain		Interactors
Yeast EDE1		AEP1, AKL1, APL1, APM2, ARE1, ATG26, ATP25, BUD3, BZZ1, CHC1, CLC1, CLN2, CMD1, CMK1, CMK2, COP1, CYM1, CYS4, DUF1, ECM25, ECM29, EDE1, END3, ENO1, ENT1, ESA1, ETP1, FRK1, HRR25, IMG1, IMG2, IST2, LAS17, LSB3, MET10, MHR1, MRM1, MRP20, MRP49, MRP7, MRPL1, MRPL10, MRPL13, MRPL15, MRPL16, MRPL17, MRPL20, MRPL22, MRPL23, MRPL24, MRPL27, MRPL28, MRPL3, MRPL35, MRPL36, MRPL39, MRPL4, MRPL44, MRPL6, MRPL7, MRPL8, MRPL9, MSS51, NAF1, NUP42, OYE2, PAL1, PHO85, PKC1, PRK1, PSH1, PXA1, RAD53, RET2, ROM2, RPA34, RPL11A, RPL12A, RPL13B, RPL16B, RPL19A, RPL28, RPL2A, RPL35A, RPL38, RPL6B, RPL9A, RPN11, RPS11A, RPS14B, RPS31, RPS6A, RSC8, RTC6, SEC26, SEC28, SEC9, SGM1, SHQ1, SLA1, SLA2, SMT3, SOV1, SPC72, SPR3, SRO9, SSC1, STM1, SWE1, SYP1, TRM3, UBI4, UTP20, YAP1802, YCK1, YML6, YPT7
Worm ITSN-1	Non-SH3	ALX-1, APA-2, APB-1, C08F8.2, C11E4.6, C12D8.1, CED-6, CHC-1, CLIC-1, DAB-1, DPY-23, DYN-1, EHS-1, EPN-1, EPS-8, F10D7.5, F10E9.3, F41H10.3, F46H5.7, F56D3.1, F59E12.9, F59G1.8, FLN-1, HCP-2, HGRS-1, HPK-1, HPO-34, IFA-1, ITSN-1, KEL-8, LEV-11, LFI-1, LIN-65, MSP-10, MSP-113, NURF-1, PCF-11, PQN-87, RME-4, SAP-49, SAS-5, SEC-31, STAM-1, SUMV-1, TAG-163, TFG-1, UBQL-1, UNC-11, W09D10.1, WWP-1, Y106G6D.7, Y44E3A.4
	SH3 #1	C06A6.2, C36C9.1, CSN-5, FIGL-1, LST-1, MAB-10, SQV-4, Y43F8C.6
	SH3 #2	UNC-26

	SH3 #4	B0286.3, NUM-1, Y57A10A.1, Y57G11C.22
	SH3 #5	B0286.3, Y77E11A.2
Human ITSN1	Non-SH3	AGFG1, AGFG2, AP2B1, ARF6, ARFIP2, ARHGAP31, BECN1, CCNO, CDC42, CEP85L, CLIP2, CLTC, CSNK2B, CYTH1, DES, DISC1, DLGAP1, EEF1A1, EPHB2, EPN2, EPS15, EPS15L1, FCHSD2, FNBP1, FNBP4, GCC1, GOLGA5, GP6, HIP1, INPPL1, ITSN1, ITSN2, KHDRBS1, KIF16B, KIF5A, LMO4, MAPK6, MAPK8IP2, MRPL20, MTUS2, PACSIN3, PAK3, PDCD6IP, PDE4D, PFDN5, PHGDH, PICALM, PIK3AP1, PIK3R1, PK, PLK1, PPFIA2, PPL, PREX1, RAB11FIP2, RAB5A, RABEP1, RAI14, RNF40, RPS6KA5, SCAMP1, SCOC, SF3B4, SGIP1, SH3GL2, SH3KBP1, SMARCC2, SMNDC1, SNAP23, SNAP25, SNX5, SPDL1, STON2, SYNJ2, TK1, TRIM8, TSG101, UBE2K, UNC119, WAS, WASL, WBP11, ZFPM2
	SH3 #1	ADAM12, ANK3, ARAP1, CBL, CBLB, CBLC, CNTN2, DAB2, DAG1, DOCK3, DOCK4, FAM162B, FGD5, GAREM, GAREML, HCN2, HERC1, IL31RA, LATS1, MYO15A, NRG1, OBSCN, PDZD8, PIK3C2B, REPS1, REPS2, RHOU, RUSC1, SDC3, SH3D19, SH3PXD2B, SIRT1, SOS1, SOS2, SPRY2, SYNJ1, TNK2
	SH3 #2	RICTOR
	SH3 #3	ADAM8, ANKRD55, ARHGAP32, ARHGEF5, ASAP1, ASAP2, BCAR1, CACNA1D, CACNA1E, CACNA1F, CAMKK1, CAMSAP1, CBL, CBLC, CELSR3, CEP170, CEP170B, COL12A1, DAG1, DNM1, DNM3, DOCK3, DOCK4, DOCK5, FNDC1, GAREM, GAREML, HCN4, HERC1, KCNA5, NRG1, OBSCN, OBSL1, PCLO, PIK3C2B, PIK3R2, PKD1, PTK2, PTPRN2, REPS1, REPS2, RET, RIMS1, RIMS2, SCN2A, SH3PXD2B, SOS1, SOS2, SPRY2, ST5, TEC, TNK2, TRIM67
	SH3 #5	AMPH, ARAP1, C1ORF168, C21ORF58, C2CD2, CASKIN1, CBL, CBLB, CDC42EP2, CNTN2, DAB2, DAG1, DLG5, DNM1, DNM2, DNM3, DOCK4, DSCAML1, FAM43A, FAT1, FGD5, GAREM, GAREML, HCN3, HERC1, KNDC1, MAP4K4, MYO15A, NCKIPSD, OBSCN, PDZD8, PEAK1, PHLDB1, PIK3C2B, PKN3, PRKCDBP, PTK2, PTK2B, REPS1, RHOU, RIMS1, RIN1, ROBO2, RUSC1, SETD5, SH3D19, SH3PXD2B, SHANK2, SIPA1L3, SOS1, SOS2, SPRY2, SYNJ1, TNK2, TP73, TTBK2, WDR44, ZNF474
Human ITSN2	Non-SH3	AGT, AHDC1, AMPH, ANKRD17, BCCIP, CCDC88C, CHIC2, CPSF6, DLGAP1, DST, EGFR, EIF3A, EIF3B, EIF3C, EIF3E, EIF3G, EIF3H, EIF3I, EIF3L, EIF4A3, EPS15, EPS15L1, ERC1, FASLG, FCHSD1, FCHSD2, FNBP4, GOLGA2, GOLGA8A, GOLGB1, GPNMB, HNRNPK, HOOK2, ITPKA, ITSN1, ITSN2, KCTD10, KDM1A, KHDRBS1, KIAA1549, KXD1, LARP6, LSM8, LTBP4, LUC7L3, MAP4K3, MBNL1, MCRS1, MEGF10, NBR1, NR2C2, PDCD6IP, PDE4DIP, PIK3AP1, PPP1CC, PSEN1, PTN, RABEP1,

		RBMX, RNF20, ROCK1, RUFY1, SEMA6A, SH3GL1, SH3KBP1, SNAP29, SRGAP2, STX4, SYN1, SYNJ2, TACC1, TBL3, TMF1, TRIO, WAS, WASF2, WASL, WIPF2, YLPM1, YTHDF1
	SH3 #1	ABR, ARAP1, CBL, CBLB, DNM2, DOCK4, DSCAML1, DYRK2, FGD5, GAREML, GDAP2, HERC1, LATS1, MYO15A, NCF1, NHSL2, OBSCN, PDZD8, PHLDB1, PIK3C2B, PRRC2B, RBM33, REPS1, RIMS1, RUSC1, SIPA1L3, SIRT1, SOS1, SOS2, TNK2
	SH3 #2	ABL2, CUX1, PRKD2, SYNE2
	SH3 #3	ANKRD55, APC, ARAP1, ARAP2, ARHGAP21, ARHGAP30, ARHGEF5, ASAP1, ASAP2, CBL, CBLB, CDK12, CEP104, CEP170, DENND1C, DOCK4, DOCK5, FGD1, HERC1, LRRC27, MYO15A, OBSCN, PIK3R1, PTK2, RIMS1, RIMS2, SOS2, SYNE2, TAGAP, TNIK, WIPF1
	SH3 #5	AAK1, ARAP1, ARHGEF15, ASAP1, ASAP2, BIN1, C2CD2, CASKIN1, CBL, CBLB, CBLC, DEPDC5, DOCK4, DOCK5, DSCAML1, EVL, FGD5, FSIP2, GAREML, HCLS1, HERC1, INPPL1, IQSEC2, KNDC1, MAP4K4, MICAL1, MYO15A, NCF1, NOXO1, OBSCN, OPHN1, PHLDB1, PIK3C2B, PKN3, PRR27, PRRC2B, PTPN12, PTPN22, RAPH1, RBM26, REPS1, RIMS1, RIN1, RIN3, ROBO1, RUSC1, SH3D19, SHANK1, SIK1, SIPA1L3, SOS1, SOS2, SPATA2L, TNK2, TTBK2, UTRN, WDR44

Table B-1 – The protein interactors for human ITSN1 and ITSN2, as well as their yeast and worm orthologs. Using SH3 binding data, we identify various PPIs known to be mediated by those SH3 domains and split out their binding partners. Protein interactors not known to bind to an SH3 domain are all grouped under "Non-SH3". While the worm SH3 binding targets are drawn from experimental data, the human SH3 binding targets are drawn from predicted data; thus, the total number of interactors assigned to human ITSN1 and human ITSN2 in this table exceed the values presented in Table **3-5**.