PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/94194

Please be advised that this information was generated on 2017-11-27 and may be subject to change.
Evolutionary analysis in PPI networks and applications

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 15 oktober 2012
om 15:30 uur precies

door

Pavol Jancura
geboren op 1 oktober 1983
te Svidník, Slowakije.
Promotor:
Prof. dr. Tom Heskes

Copromotor:
Dr. Elena Marchiori

Manuscriptcommissie:
Prof. dr. ir. Theo P. van der Weide
Prof. dr. Jaap Heringa (Vrije Universiteit Amsterdam)
Dr. Franca Fraternali (King’s College London, United Kingdom)

SIKS Dissertation Series No. 2012-34
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Copyright © 2012 Pavol Jancura, Nijmegen, The Netherlands
Printed by Ipskamp Drukkers, Nijmegen
Acknowledgements

The completion of this thesis would not be possible without the proper supervision, collaboration, and support of many people I met in my professional or personal life to whom I would like to give my acknowledgements.

First of all, I would like to thank and express my deep gratitude to my main supervisor Elena Marchiori. She not only provided me the regular and professional guidance, but always filled me with enthusiasm, joy, and positive attitude towards my work. I would like to thank to my promotor Tom Heskes that I could join his group and conduct the research summarized in this thesis. I am grateful for the time he always found to offer a constructive discussion and counselling when I needed it.

I am thankful to Jaap Heringa for the opportunity to start my doctoral research at Centre for Integrative Bioinformatics at Vrije University after finishing my master studies and that I could be part of his group before moving to Nijmegen. My thanks belong to Martijn Huynen for his critical view and biological perspective on my work and that I could spent a part of my doctoral studies in his group at Centre for Molecular and Biomolecular Informatics at Radboud University Nijmegen Medical Centre.

I would like to acknowledge my collaborators Eleftheria Mavridou and Enrique Carrillo-de Santa Pau for their time, discussions, and contributions that improved my work and for everything I could learn from them. I thank to Beatriz Pontes for her technical and programming support.

I would like to offer special thanks to my colleagues Ali Bahramisharif, Saiden Abbas, Fabio Gori, Adriana Birlutiu, Rasa Jurgenelaite, Botond Cseke, Tom Claassen, Perry Groot, Marcel van Gerven, Tjeerd Dijkstra,
Janos Sarbo, Evgeni Tsivtsivadze, Dimitris Mavroeidis, Joris Mooij, Twan van Laarhoven, Wout Megchelenbrink, Max Hinne, Saskia Koldijk, Maya Sappelli, and Daniel Kühlwein for the time we spent together and for sharing their valuable insights and comments to my work. I am particularly grateful to Twan van Laarhoven for helping me with the Dutch translation of the summary of this thesis. I would like to convey thanks to Nicole Messink for her administrative help and assistance.

I wish to thank to all my friends for their support and fellowship including Cezary Kaliszyk, Łukasz Chmielewski, Michal Juríček, Dáša Lauková and Ján Lauko, Georgeta Igra, Olha Shkaravska and Luminita Moruz. I thank to Marc Jupin, the Poos family, the van Rossum family, the Swaenberg family, Ft. Ewald Kamphuis and Ft. Cyrus van Vught, and all other people of the Maria Geboorte Church community for their friendship that enriched and deepen my life.

My thanks also belong to my parents, Anna and Ján, for everything they have done for me. I thank to my sisters, Kristína and Jana, and my brother Gregor for their words of support and encouragement.

Last but not least, I would like to express my gratitude and thanks to my beloved wife Patrycja and my daughter Antonina for their love, tenderness, and patience, while writing this thesis.
## Contents

Acknowledgements iii

1 Introduction 1
   1.1 The thesis outline 5

2 A survey on evolutionary analysis in PPI networks 7
   2.1 Introduction 8
   2.2 Unravelling the relations between evolution and PPI network structure 10
      2.2.1 Relation between a single protein in a PPI network and evolution 11
      2.2.2 Higher-order structures in a PPI network and evolution 19
   2.3 Using evolutionary information for knowledge discovery in PPI networks 26
      2.3.1 Predicting protein interaction: Interologs 27
      2.3.2 Pairwise protein network alignment 30
      2.3.3 Multiple protein network alignment 38
      2.3.4 Applications and future developments 41

3 Quantifying the orthology signal in a PPI network at a protein complex level 45
   3.1 Introduction 46
   3.2 Methods 49
# Contents

3.2.1 Data ................................................. 49
3.2.2 Quantifying orthology signal ......................... 51
3.3 Results and discussion ............................... 57
  3.3.1 Generating the cluster classes ...................... 57
  3.3.2 Identifying orthology signal at protein complex level 59
  3.3.3 On orthology-related functions ..................... 61
  3.3.4 On orthology-related complexes and novel predictions 63
3.4 Conclusions ........................................... 69

4 Dividing protein interaction networks for detection of conserved complexes 71
  4.1 Introduction ......................................... 72
  4.2 Related work ........................................ 74
  4.3 Graph theoretic background .......................... 75
  4.4 Divide algorithm ..................................... 78
  4.5 Experimental analysis ............................... 81
  4.6 Divide generates sub-graphs covering ”true” protein conserved complexes ......................... 83
  4.7 Comparison of Divide sub-graphs with predicted conserved complexes ....................... 87
  4.8 Conclusions .......................................... 88

5 Modular alignment of protein interaction networks 91
  5.1 Introduction ......................................... 92
  5.2 Related work ........................................ 93
  5.3 Modular network alignment with Divide ............... 96
    5.3.1 DivAfull ......................................... 97
    5.3.2 Divide and MNAligner ............................ 99
    5.3.3 Handling redundant alignments and complexes 100
  5.4 Evaluation criteria for conserved complexes .......... 103
  5.5 Results ............................................. 106
    5.5.1 Application of Divide ........................... 107
    5.5.2 DivAfull and MaWish ............................ 107
    5.5.3 Divide and MNAligner ............................ 110
Chapter 1

Introduction

Successful and wide applications of computer science and applied mathematics in life sciences led to the emergence of new interdisciplinary research fields in biology, such as bioinformatics and computational biology. As a result, tools and methods to process or to produce large amount of data have been introduced as well as computational approaches to directly address and test biological hypotheses. The broad development and use of bioinformatics tools and computational biology approaches give great prospects to systems biology and computational translational medicine.

One of the main achievements in bioinformatics consists of the development of methods for sequence comparison. Since the advent of global and local sequence alignment (Needleman and Wunsch, 1970, Smith and Waterman, 1981), methods for this task have advanced immensely in their time efficiency and accuracy, thus enabling sequence alignment of multiple species and large-scale analysis of genomic data. The development of comparative tools and the constantly better access to the increasing number of sequenced genomes have elevated the importance and significance of another, relatively young, interdisciplinary field, comparative genomics. Comparative genomics provides an evolutionary perspective on the interpretation and analysis of genomic data across species. Its aim is to identify genome structures that are shared or different in distinct organisms. Es-
establishing such inter-genomic maps at various organizational levels enable to identify and study evolutionary traits and processes that shape genome evolution. Moreover, comparative genomic methods offer powerful means for cross-species inference of biological information and functional genome annotation.

The basic premise of comparative genomics is that those genomic features responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences, while those elements responsible for differences among species should be divergent (Hardison, 2003). As a result, one may identify similar genes or proteins, so-called homologs, by means of sequence alignments methods on nucleotide or amino-acid sequence level, respectively. However, the presence of homologous genes can have various origins due to multitude of possible evolutionary events. There are two fundamental types of homologous genes, orthologs and paralogs, where the first type of homologs evolved by speciation through a vertical descent from a single ancestral gene and the latter by duplication process (Fitch, 1970). Other evolutionary scenarios, such as lateral gene transfer, gene loss or fusion/fission events, are also possible yielding the consecutive classification of orthologs and paralogs into other subtypes (Koonin, 2005). It is, therefore, difficult to establish proper orthology between genes and proteins, and hence simplified, computational definitions of orthologs and paralogs are used, especially for the purpose of identifying classes of functionally equivalent genes (Koonin, 2005, Kuzniar et al., 2008).

As the genome encodes all of the biological information needed for development and functioning of an organism, the complexity of its evolution influences every level of a living system. Thereby, one can presume the presence of evolutionary traits also between molecular interactions and metabolic processes. The possibility to study these characteristics comes with the technological advancement that allows to generate data on proteomic and genetic interactions and metabolic pathways.

In particular, high-throughput techniques, such as the yeast two-hybrid assays (Y2H) (Chien et al., 1991, Fields and Song, 1989), tandem affinity purification (TAP) (Rigaut et al., 1999) or co-immunoprecipitation fol-
allowed by mass spectrometry (AP/MS) (Hartman et al., 2001, Ho et al., 2002), have enabled the accumulation of data on physical protein-protein interactions (PPI) and the construction of publicly available interactome databases, as for example IntAct (Kerrien et al., 2012), MINT (Licata et al., 2012) or BioGRID (Stark et al., 2011). Such interactome data can be modelled as large-sized graphs, called PPI networks, where proteins are represented as nodes of a network and an edge between nodes corresponds to an interaction between respective proteins (see Figure 1.1).

These biological networks, in general, can exhibit many properties and inner structures, which, in case of PPI networks, may relate to various biological functions and organizational levels. The characteristics of PPI networks can also be evolutionary driven. Therefore, network analysis offers a powerful approach to understanding the biological organization and the function of cellular components in living organism and it can bring new insights into evolutionary principles (Vespignani, 2003). In addition to physical interactions, a network model can be further strengthen by integrating various experimental or computational evidences on the functional inter-play between proteins yielding a PPI network of functional associations or functional PPI network (Bebek et al., 2012, Srinivasan et al., 2007).

To trace the intrinsic evolutionary imprints in the structure of interaction networks one can either study a single PPI network and how its topology is constrained by evolution or compare multiple networks of different species to identify their homologous regions and investigate their conservation and divergence at present time (Jancura and Marchiori, 2012). When a single network is considered, researchers focus on the link between genomic evolution and the connectivity properties of proteins and interactions to learn about network evolution. In addition, as interactions are the result of the fact that proteins often do not act alone but form complexes or modules to perform a certain biological task, the evolutionary pressures on modularity structure of a network are also highly investigated. Understanding the evolutionary origin of functional modules may help to explain their spatio-temporal characteristics as well as their functional relevance in terms of dispensability or essentiality of proteins (Campillos et al., 2006).

The second perspective considers comparison of multiple networks across
Figure 1.1: A layout of the highly-curated yeast PPI network data published in (Batada et al., 2007) as visualized by Cytoscape software (Smoot et al., 2011).

distinct species and its motivations are analogous to genome comparison. Network comparison or network alignment aims to address similar biological questions as sequence alignment. The main task is to identify proteins, protein interactions, or protein complexes/modules that are likely to possess equivalent functions across species (Sharan and Ideker, 2006). As a result, the established homologies may be used to predict new functional
information about proteins and interactions that are poorly characterized. Moreover, it can bring a deeper understanding into the evolution of proteins, networks and whole species as shown in (Sharan and Ideker, 2006). For example, knowing putative orthologous protein complexes can help to build species’ phylogeny and to study protein complex evolution over the evolutionary tree (Erten et al., 2009, Yosef et al., 2009).

1.1 The thesis outline

The work in this thesis compiles specific contributions to the field of evolutionary analysis in PPI networks. In particular, we study the presence of evolutionary protein complexes in a species’ physical interactome produced by high-throughput experiments. The thesis is organized as follows.

In the next chapter (Chapter 2) we provide an extensive literature survey on evolutionary analysis in PPI networks. This gives the reader a broad overview on key issues tackled in the field and on the various approaches employed. It presents literature evidence on some hypotheses made about network and protein module evolution as well as main methodological concepts for comparing the interactome. The chapter also discusses successful applications of network alignment methods and their future perspectives.

Chapter 3 analyses the effect of the evolutionary signal induced by the presence of evolutionary conserved proteins on the detection of protein complexes in a single PPI network. Although the study considers interactome of only one species, the proposed methodology reveals that the protein complexes driven by the signal are clearly differentiated from the complexes obtained with random or no bias. Moreover, the method proves that the evolutionary signal is quantifiable with biological functions of the evolutionary-driven complexes.

Next, Chapter 4 presents an algorithm that aims to decompose the network of a single species into smaller network regions such that these regions contain evolutionary conserved complexes. In contrast to evolutionary-driven complexes in Chapter 3, evolutionary conserved complexes have a stricter definition because conserved complexes must have their homologous
counterparts in the network of another species, while evolutionary-driven complexes need not satisfy this constrain. The effectiveness of the method is demonstrated by extensive experimental analyses, which in particular show that the evolutionary conservation of protein complexes in PPI networks can be studied locally.

In Chapter 5 we further exploit the divisive algorithm proposed in Chapter 4. Specifically, we perform network alignments of two species in a modular fashion by preprocessing each PPI network prior to their alignment with that algorithm. These applications verify a proof-of-principle for modular approach to network alignment and the final results substantiate the use of the divisive algorithm for improving the performance of state-of-the-art alignment methods.
Chapter 2

A survey on evolutionary analysis in PPI networks

The possibility of measuring protein evolution by computational means and an emergence of protein interaction data have led to new perspectives of research in system biology in recent years. Particularly, many researchers have contributed to the field analysing the presence of evolution in PPI networks and protein interaction modules as well as providing methods and tools for comparing various networks to identify their evolutionary common building blocks. However, a comprehensive survey on the topic has been missing in literature. Thus, we present an extensive overview on evolutionary analysis in PPI networks and its applications.¹

¹This chapter is based on the following work:
Chapter 2. A survey on evolutionary analysis in PPI networks

2.1 Introduction

The analysis and application of the evolutionary information, as measured by means of the conservation of protein sequences, using protein-protein interaction (PPI) networks, has become one of the central research areas in systems biology from the last decade. It provides a promising approach for better understanding the evolution of living systems, for inferring relevant biological information about proteins, and for creating powerful protein interaction and function prediction tools. The aim of this survey is to give a general overview of the relevant literature and advances in the analysis and application of evolution in PPI networks. Due to the broad scope and vast literature on this subject, the present overview will focus on a representative selection of research directions and state-of-the-art methods to be used as a solid knowledge background for guiding the development of new hypothesis and methods aiming at the extraction and exploitation of evolutionary information in PPI networks.

This survey consists of two main parts (see Figure 2.1). The first part (Section 2.2) deals with research works concerning the relation between evolution and the topological structures of a PPI network, in particular trying to discover and assess the evidence of such a relation and its strength at different granularity levels. Specifically, we consider works analysing evolution at the single protein level as well as at the level of a collection of proteins present in a PPI network. The second part of this survey (Section 2.3) describes works analysing how such evolutionary evidence can be exploited for knowledge discovery, in particular for inferring relevant biological information, such as protein interaction prediction and the discovery of functional modules conserved across multiple species.

The main terms and concepts underlying protein interaction and evolution which are used throughout the survey are summarized in the sequel. In general, a protein-protein interaction can represent different types of relations, such as a true physical bond or a functional interplay between proteins. Here, if not explicitly stated, a PPI represents a physical protein interaction as detected by experimental methods, such as yeast two-hybrid screening, co-immunoprecipitation or tandem affinity purification.
2.1. Introduction

Evolutionary information & PPI networks

2.2 Unravelling the relations between evolution and network structure in PPI networks

2.2.1 A single protein

Central vs Non-central

Party vs Date hubs

Supporting node connectivity relevance for protein evolution

2.2.2 Higher-order network structures

Protein-protein interaction Protein modules

2.3 Using evolutionary information for knowledge discovery in PPI networks

2.3.1 Interologs

2.3.2 Pairwise Network Alignment

Local: pathways Local: complexes Global

2.3.3 Multiple Network Alignment

2.3.4 Applications & future developments

Figure 2.1: The structure of the survey.

Two proteins are called homologous if they share high sequence similarity. There are two main types of homologous proteins: orthologous and paralogous. Here, for simplicity, we consider a protein pair to be orthologous if the proteins of the pair are from different species. We refer to the proteins of an orthologous pair as orthologs. Analogously, a protein pair is considered to be paralogous if its proteins belong to the same species, in this case their proteins are called paralogs. A general assumption is that the proteins of an orthologous pair originated from a common ancestor, having been separated in evolutionary time only by a speciation event, while paralogous proteins are the product of gene duplication without speciation. The concept of orthology can be directly extended to more than two species,
Chapter 2. *A survey on evolutionary analysis in PPI networks*

where one can consider clusters of orthologous proteins containing at least one protein of each species.

2.2 Unravelling the relations between evolution and PPI network structure

We begin with a summary of those studies that involve the analysis of evolutionary information in a single PPI network. One can divide these works into the following two main groups. The first group studies evolutionary conservation with respect to topological properties of a PPI network. The second one primarily investigates the role of evolution with respect to the functional modules present in a PPI network.

The aim of the first group of studies is to describe how the topology of a single PPI network reflects the evolutionary signal present in the proteins it contains. This evolutionary signal is represented by the set of orthologs and it is retrieved with respect to a different species. Specifically, given a PPI network of the species to be investigated and a set of proteins of a distinct species, those proteins of the network being a part of orthologous pairs or clusters (resulting from a sequence comparison of proteins of the two or multiple species respectively) are considered to be source of the evolutionary or orthology signal in the network. Then, having established the orthology relationship between proteins of the two or multiple species, one can estimate the evolutionary rate or distance of aligned protein sequences (see e.g. Grishin, 1995). The higher the rate, the faster is considered the evolution of proteins. Consequently, proteins which evolve slowly are well-conserved and a little or none change to them can be observed throughout the evolution. Other protein evolutionary measures have been considered, as propensity for gene loss, evolutionary excess retention or protein age (see Table 2.1).
2.2. Unravelling the relations between evolution and PPI network structure

2.2.1 Relation between a single protein in a PPI network and evolution

Various features of a PPI network topology can be investigated with respect to evolutionary information; the first and simplest ones are measures acting on the single nodes of the network. One can associate with a node different topological measures which estimate the relative relevance of the node within the network, here called centrality or connectivity of a node.

A basic centrality measure of a node is its degree. The degree of a node is the number of edges containing the node or, in terms of a PPI network, it is the number of proteins with which the protein represented by the node in the network interacts. It has been observed that a protein degree distribution of PPI networks follows a power law and thus PPI networks fall into a class of scale-free networks (see e.g. Jeong et al., 2001, Park et al., 2001, Wagner, 2001). Scale-free networks have a few highly connected nodes, called hubs, and numerous less connected nodes, which mostly interact only with one or two nodes.

<table>
<thead>
<tr>
<th>Type of evolutionary measure</th>
<th>Evolutionary measure</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolutionary conservation</td>
<td>Evolutionary Rate</td>
<td>Grishin (1995)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yang and Nielsen (2000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wall et al. (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anisimova and Kosiol (2009)</td>
</tr>
<tr>
<td></td>
<td>Evolutionary Excess Retention</td>
<td>Wuchty (2004)</td>
</tr>
<tr>
<td></td>
<td>Phyletic Retention</td>
<td>Fang et al. (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chen and Xu (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gustafson et al. (2006)</td>
</tr>
<tr>
<td></td>
<td>Protein Age Group</td>
<td>Ekman et al. (2006),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kim and Marcotte (2008)</td>
</tr>
</tbody>
</table>

Table 2.1: Measures of evolutionary signal at protein level
Chapter 2. A survey on evolutionary analysis in PPI networks

Essentiality, centrality and conservation of a protein

As a decade ago large protein physical interaction data were not yet available, researchers mainly focussed on the study of the correlation between importance of a protein function for a living cell (essentiality, dispensability) and its evolutionary conservation rate. The generally accepted premise is that essential genes or proteins should evolve at slower rates than non-essential ones (Kimura, 1983, Kimura and Ohta, 1974, Wilson et al., 1977). Although empirical studies have cast doubts on the validity of this hypothesis (see Table 2.2), in the end the vast majority and late evidences favour the existence of correlation between gene essentiality or dispensability and evolutionary conservation (see Table 2.3). In particular, as recently stated by Wang and Zhang (2009), the correlation remains weak yet still conveniently sufficient for practical use.

After the growth of protein interaction data, also the correlation between essentiality and centrality, and evolutionary conservation and centrality started to be investigated. At first the centrality-essentiality relationship was primarily studied by examining the degree of a node, proving the existence of the correlation (see e.g. Fraser et al., 2003, 2002, Hahn and Kern, 2005, Jeong et al., 2001, Krylov et al., 2003). However Coulomb et al. (2005) showed no correlation between essentiality and centrality, where centrality was assessed not only by the degree but also by higher order centrality measures, namely average neighbours’ degree of a node and clustering coefficient of a node, suggesting that the correlation centrality-essentiality could be an artefact of the dataset. These findings were later supported by Gandhi et al. (2006) who considered a set of PPI networks and also did not observe any significant relationship between a node degree and the essentiality of the corresponding protein. Interestingly, Coulomb et al. (2005) did not test other centrality measures as betweenness and closeness, which showed a higher correlation with essentiality than just the simple degree (Hahn and Kern, 2005). Nevertheless, Batada et al. (2006a) reaffirmed the existence of the correlation between the node degree and essentiality taking into account Coulomb et al.’s concerns. However, Yu et al. (2008) again disputed the correlation using the compilation of a yeast high quality PPI
2.2. Unravelling the relations between evolution and PPI network structure

data. Results contradicting the work of Yu et al. appeared in two consecutive studies by Park and Kim (2009) and Pang et al. (2010b). The first study (Park and Kim, 2009) considered also other centrality measures than just the degree of a node. As a result, the correlation could be successfully revealed, whereas the highest correlation was observed with measures based on betweenness and closeness, similarly to Hahn and Kern (2005). In the other study (Pang et al., 2010b) the newer, updated yeast PPI dataset was used and the correlation between degree of a node and its (protein) essentiality could be detected.

Although, the above works support that there is a connection between topological position of a node and functional importance, it seems one cannot explain this centrality-lethality rule just by the degree distribution (He and Zhang, 2006, Zotenko et al., 2008). This seems to be in accordance with the analysis conducted in (Lin et al., 2007) showing that protein domain complexity is not the single determinant of protein essentiality and that there is a correlation between the number of protein domains and the number of interactions (Schuster-Bockler and Bateman, 2007). In addition, Kafri et al. (2008) showed that highly connected essential proteins tend to have duplicates which can compensate their deletion thus decreasing the deleterious effect of their removal, a phenomenon that could possibly explain the findings that genes with no duplicates are more likely to be essential (Giaever et al., 2002). Therefore higher order topological features appear to be more appropriate for capturing gene essentiality, especially those based on node-betweenness and node-closeness (Hahn and Kern, 2005, Park and Kim, 2009, Yu et al., 2007), which are believed to estimate better the local connectivity or centrality of a node within the network. Moreover, these features also relate with gene expression (Krylov et al., 2003, Pang et al., 2010b, Yu et al., 2007).

We consider now works that analyse the correlation between evolution and centrality. Also in this case the two main features used to estimate this correlation are the degree of a node and the evolutionary rate. At first, it was hypothesized that proteins with a higher degree should evolve slower (Fraser et al., 2002). A main criticism to this hypothesis was based on the fact that the analysis conducted in (Fraser et al., 2002) did not
take into account the presence of a possible bias and of noise in data obtained from high-throughput experiments (Bloom and Adami, 2003, 2004, Jordan et al., 2003a,b). Nevertheless Fraser et al. (2003), Fraser and Hirsh (2004) and Lemos et al. (2005) could confirm the existence of such correlation by taking into account these objections. Kim et al. (2007) also confirmed interconnection between centrality, essentiality and conservation and showed that peripheral proteins of the PPI network are under positive selection for species adaptation. Moreover, the link between the connectivity of a node and its evolutionary history was further substantiated by works studying the correlation between node degree and other evolutionary measures such as propensity for gene loss (Krylov et al., 2003), evolutionary excess retention (Wuchty, 2004) and protein age (Ekman et al., 2006, Kunin et al., 2004). However Batada et al. (2006a) again pointed to a lack of evidence for a significant correlation between the evolutionary rate and the connectivity of a node. Moreover, Makino and Gojobori (2006) classified proteins according to two criteria, clustering coefficient of a node and protein’s multi-functionality, and showed that multi-functional proteins of sparse parts of yeast PPI network (with a low clustering coefficient) evolve at the slowest rate regardless of the degrees of the connectivity. This suggests that clustering coefficient is a better descriptor of protein evolution within the global network of protein interactions.

A possible explanation for these conflicting results was proposed by Saeed and Deane (2006) who showed that the strength and significance of the correlation between evolution and centrality varies depending upon the type of PPI data used. Also Saeed and Deane (2006) found that more accurate datasets demonstrate stronger correlations between connectivity and evolutionary rate than less accurate datasets. Another reason may be the existence of two distinct types of highly connected nodes, so-called party and date hubs, which appear to satisfy different evolutionary constraints.

Evolution of party and date hubs

Specifically, Han et al. (2004) observed a bimodal distribution of average Pearson correlation coefficients between the expression profiles of proteins
2.2. Unravelling the relations between evolution and PPI network structure

<table>
<thead>
<tr>
<th>Essentiality-Evolution</th>
<th>Essentiality-Centrality</th>
<th>Evolution-Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hurst and Smith (1999)</td>
<td>Coulomb et al. (2005)</td>
<td>Jordan et al. (2003a,b)</td>
</tr>
<tr>
<td>Rocha and Danchin (2004)</td>
<td></td>
<td>Batada et al. (2006a)</td>
</tr>
<tr>
<td>Drummond et al. (2006)</td>
<td></td>
<td>Drummond et al. (2006)</td>
</tr>
</tbody>
</table>

Table 2.2: Literature evidence on no correlation between centrality, evolution and essentiality of proteins

and its interacting partners. This yielded a classification of hubs into party hubs, having similar co-expression profiles with their neighbours, and date hubs, having different co-expression profiles with their neighbours. As a consequence, party hubs tend to interact simultaneously (“permanently”) with their partners and to connect proteins within functional modules while date hubs tend to interact with different partners at different time/space (“transiently”) and to bridge different modules. Thus, one may also refer to party hubs as intramodule and to date hubs as intermodule (Fraser, 2005).

Fraser (2005) was the first to investigate the difference in evolution between date and party hubs and found that party hubs are highly evolutionary constrained, whereas date hubs are more evolutionary labile. This is clearly in accordance with findings of Mintseris and Weng (2005) who argued that residues in the interfaces of permanent protein interactions tend to evolve at a relatively slower rate, allowing them to co-evolve with their interacting partners, in contrast to the plasticity inherent in transient interactions, which leads to an increased rate of substitution for the interface residues and leaves little or no evidence of correlated mutations across the interface. The work of Fraser (2005) was, in addition, later corroborated by Bertin et al. (2007). Examining three dimensional properties of proteins also supported this hypothesis, as multi-interface hubs were found to be more evolutionary conserved and essential as well as more likely to correspond to party hubs (Kim et al., 2006). Defining singlish- and multi-Motif hubs further substantiated these findings, because multi-Motif hubs were found to be more evolutionary conserved, more essential and to correlate with multi-interface hubs (Aragues et al., 2007). In addition, other features
Chapter 2. A survey on evolutionary analysis in PPI networks

<table>
<thead>
<tr>
<th>Essentiality-Evolution</th>
<th>Essentiality-Centrality</th>
<th>Evolution-Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castillo-Davis and Hartl (2003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang and He (2005)</td>
<td></td>
<td>Lemos et al. (2005)</td>
</tr>
<tr>
<td>Wall et al. (2005)</td>
<td>Batada et al. (2006a)</td>
<td>Butland et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>Zotenko et al. (2008)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Park and Kim (2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acencio and Lemke (2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manimaran et al. (2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hwang et al. (2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ning et al. (2010)</td>
<td></td>
</tr>
<tr>
<td>Dotsch et al. (2010)</td>
<td>Pang et al. (2010b)</td>
<td>Pang et al. (2010a)</td>
</tr>
<tr>
<td>Theis et al. (2011)</td>
<td>Theis et al. (2011)</td>
<td>Theis et al. (2011)</td>
</tr>
</tbody>
</table>

Table 2.3: Literature evidence on correlation between centrality, evolution and essentiality of proteins
2.2. Unravelling the relations between evolution and PPI network structure

as orderness of regions in protein sequences and the solvent accessibility of the amino acid residues was shown to be different between party and date hubs and to contribute in the lowering of the evolutionary rate of party hubs (Kahali et al., 2009). Recently, Mirzarezaee et al. (2010) applied feature selection methods and machine learning techniques to predict party and date hubs based on a set of different biological characteristics including amino acid sequences, domain contents, repeated domains, functional categories, biological processes, cellular compartments, etc.

However, other researchers disputed not only the evolutionary differences between party and date hubs but the existence of hub types as such (Agarwal et al., 2010, Batada et al., 2006b, 2007). Indeed, some datasets do not exhibit clear or robust bimodal distribution of hubs’ gene co-expression profiles (Agarwal et al., 2010, Brown and Jurisica, 2005, Ekman et al., 2006, Salwinski et al., 2004) and in some cases there is even a complete lack of bimodality (Batada et al., 2006b, 2007). Therefore, Pang et al. (2010a) argue that the average Pearson correlation coefficient is a weak measure of whether a protein acts transiently or permanently with its interacting partners and they propose a new measure, a co-expressed protein-protein interaction degree. This measure estimates the actual number of partners with which a protein can permanently interact. One can interpret it as a degree of ‘protein party-ness’ and it offers more a continuum-like estimate of the protein’s interaction property. This seems to be in accordance with Nooren and Thornton (2003) who suggest that rather a continuum range exists between distinct types of protein interactions and that their stability very much depends on the physiological conditions and environment.

Pang et al. (2010a) firstly corroborated the results of Saeed and Deane (2006) on the correlation variations between connectivity and evolutionary rate of a protein on different datasets and then they showed that the co-expression-dependent node degree correlates significantly with the protein’s evolutionary rate irrespectively of the specific dataset used. However, their topological measure is derived by using an external source of experimental data on gene expression. The further investigation on purely topological features of a PPI network which would distinguish transient and permanent interactions, and party and date hubs could bring more insights on
how the evolutionary history of a protein is wired in its position within the network of all the protein interactions in an organism. In this perspective, network path-based measures, such as betweenness and closeness, seem to be promising (Yu et al., 2007). All the more, these measures also appear to relate to protein essentiality (Park and Kim, 2009, Yu et al., 2007) and it could clarify the link between essentiality and evolution as such. Thereafter, they could improve on the prediction of essential genes from the topology of a PPI network in combination with protein evolutionary information, such as phyletic retention (Gustafson et al., 2006), as already corroborated by several application of machine learning techniques for essential gene detection, prioritizing drug targets and determining virulence factors (Acencio and Lemke, 2009, Chen and Xu, 2005, Deng et al., 2011, Doyle et al., 2010, Gustafson et al., 2006, Hwang et al., 2009, Manimaran et al., 2009, McDermott et al., 2009).

Node connectivity is relevant for protein evolution

Since the factors relevant for protein evolution could be of a multiple character (Wolf et al., 2006), it is interesting to investigate whether protein connectivity plays a central or a more subtle role. In the latter case, the link between protein connectivity and evolution could be the results of spurious correlations due to other underlying biological processes (Bloom and Adami, 2003). In order to address this issue, the contribution of protein connectivity to protein evolutionary conservation has been also studied in an integrative way (Pal et al., 2006) using multidimensional methods such as principal component analysis (PCA) and principal component regression (PCR).

The first successful application of PCA was given by Wolf et al. (2006) on seven genome-related variables. The derived first component reflected a gene’s ‘importance’ and confirmed positive correlation between lethality, expression levels and number of protein-protein interaction which at the same time constrained protein evolution measures. Interestingly, the component also showed that the number of paralogs positively contributes to gene essentiality, which contradicts the finding of Giaever et al. (2002)
that non-duplicated genes tend to be essential. However, the study of Drummond et al. (2006) revealed by using PCR only single determinant of protein evolution, namely translational selection, which is almost entirely determined by the gene expression level, protein abundance, and codon bias. Later, Plotkin and Fraser (2007) re-examined the use of PCR method and showed noise in biological data can confound PCRs, leading to spurious conclusions. As a result, when they equalized for different amounts of noise across the predictor variables no single determinant of evolution could be found indicating that a variety of factors, including expression level, gene dispensability, and protein-protein interactions, may independently affect evolutionary rates in yeast. This observation was further substantiated by a recent study (Theis et al., 2011) where 16 genomic variables were analysed using Bayesian PCA. The study supports the evidence for the three above-discussed correlations. It also demonstrates how different definitions of paralogs may lead to different conclusions on their effect on essentiality, and thus commenting on Wolf et al.’s conflicting result (Wolf et al., 2006).

2.2.2 Higher-order structures in a PPI network and evolution

Researchers have also focused on other topological structures of a PPI network than just a node and their relation to evolutionary conservation. With increasing topological complexity we may talk about a single protein-protein interaction (an edge in PPI network), topological motifs, and protein clusters or modules as detected by their interaction density or network traffic.

Evolution and protein-protein interaction

Unlike in the case of a single protein, where various well-established methods for measuring sequence evolution are developed, to the best of our knowledge only a recent attempt has been made in order to estimate the evolutionary rate of protein-protein interaction (Qian et al., 2011). However, this study is limited to a small set of PPIs in yeasts and can not be yet
applied for large-scale studies due to the lack of data. Thus, the research has extensively focused on estimating correlated evolution of a protein pair and their functional or physical interaction (Pazos and Valencia, 2008).

It is generally assumed that proteins which co-evolve tend to participate together in a common biological function. This hypothesis is supported by many examples of functionally interacting protein families that co-evolve (see e.g. Cunningham et al., 2000, Galperin and Koonin, 2000, Goh et al., 2000, Moyle et al., 1994, van Kesteren et al., 1996). Co-evolution of proteins may be assessed at sequence level (sequence co-evolution) by correlating evolutionary rates (Clark et al., 2011), or at gene family level (gene family evolution) by correlating occurrence vectors (Kensche et al., 2008). An occurrence vector or a phylogenetic profile (phyletic pattern) (Tatusov et al., 1997) is an encoding of protein’s (homologue’s) presence or absence within a given set of species of interest (Kensche et al., 2008). In general, the methods for correlating protein evolution have been successfully applied to predict a physical or functional interaction between proteins (Clark et al., 2011, Kensche et al., 2008), where sequence co-evolution is powerful in predicting the physical interaction and phylogenetic profiling is a good indicator of functional interplay between proteins in a broader sense. Large-scale co-evolutionary maps have also been constructed and analysed for better understanding the evolution of a species and its link to protein interactions (Cordero et al., 2008, Juan et al., 2008, Karimpour-Fard et al., 2007, 2008, Tillier and Charlebois, 2009, Tuller et al., 2009). All these works suggest that the topology of PPIs should reflect the evolutionary processes behind the proteins which formed such network.

The first systematic study of linked genes and their evolutionary rates was done by Williams and Hurst (2000) who showed that the rates of linked genes are more similar than the rates of random pairs of genes. Pazos and Valencia (2001) performed the first successful large-scale prediction of physical PPIs based on sequence co-evolution by correlating phylogenetic trees. Another large-scale study by Kim et al. (2004) on domain structural data of interacting protein families also revealed their high co-evolution but also showed a high diversity in the correlation of rates of each family pair. Specifically, protein families with a greater number of domains were
shown to be more likely to co-evolve. However, Hakes et al. (2007) argued that this correlation of evolutionary rates is not responsible for the co-variation between functional residues of interacting proteins. Nevertheless, other studies have been able to predict interacting domains from co-evolving residues between domains or proteins (Jothi et al., 2006, Kann et al., 2007, 2009, Yeang and Haussler, 2007) indicating that different organisms use the same ‘building blocks’ for PPIs and that the functionality of many domain pairs in mediating protein interactions is maintained in evolution (Itzhaki et al., 2006). In addition, recently Cui et al. (2009) examined protein evolution on a human signalling network and showed that different types of interactions have different strength of constraints on protein co-evolution, in which proteins linked by physical interactions tend to co-evolve more.

Another perspective on co-evolution of interacting partners was given by Mintseris and Weng (2005), who distinguished between transient and obligate interactions. The authors concluded that obligate complexes are likely to co-evolve with their interacting partners, while transient interactions with an increased evolutionary rate show only little evidence for a correlated evolution of the interacting interfaces. This observation was later corroborated by Brown and Jurisica (2007) who analysed the presence of protein interactions across multiple species via orthology mapping and found that the greater the conservation of a protein interaction, the higher the enrichment for stable complexes. Beltrao et al. (2009) also observed that stable interactions are more conserved than transient interactions, by studying evolution of interactions involved in phosphoregulation. Finally, Zinman et al. (2011) extracted protein modules from a yeast integrated protein interaction network using various source of PPI evidence, and showed that interactions within modules were much more likely to be conserved than interactions between proteins in different modules. The results were examined for estimated evolutionary rates as well as for evolutionary retention of interactions across species.

The preference of conserved protein interactions to be placed in modular parts of a network was also observed by Wuchty et al. (2006) by extending the paradigm of protein’s connectivity and its evolutionary conservation to the connectivity of a protein-protein interaction. Specifically, they used
Chapter 2. A survey on evolutionary analysis in PPI networks

the hypergeometric clustering coefficient to estimate the interaction cohe-
siveness of the PPI’s neighbourhood and orthologous excess retention in
order to asses the evolutionary conservation of PPIs. They used the same
clustering coefficient as that given by the presence of orthologs of inter-
acting proteins in another organism and showed that PPIs with highly
clustered environment were accompanied by an elevated propensity for the
corresponding proteins to be evolutionary conserved as well as preferably
co-expressed (Wuchty et al., 2006). These findings are significant all the
more they were shown to be stable under perturbations. This propensity
of interacting proteins to be more conserved and prevalent among taxa was
later confirmed by Tillier and Charlebois (2009) who used evolutionary dis-
tances to estimate the protein’s conservation. Yet another perspective on
conservation of PPIs was given by Kim and Marcotte (2008) who classi-
fied proteins into four groups (from oldest to youngest) according their age
and found a unique interaction density pattern between different protein
age groups, where the interaction density tends to be high within the same
group and sparse between different age groups.

Evolution and modularity of PPI networks

All the evidences above that PPIs whose proteins are evolutionary corre-
lated tend to form stable complexes and to be embedded in cohesive areas
of a network topology support the premise that modularity of PPI networks
is maintained by evolutionary pressure (Vespignani, 2003). Indeed, when
examining networks solely built from sequence co-evolution, gene context
analysis or gene family evolution of completely sequenced genomes, one
may observe that these networks exhibit high modularity with clusters cor-
responding to known functional modules, thus revealing the structure of
cellular organization (Cordero et al., 2008, Tuller et al., 2009, von Mering
et al., 2003).

Regarding the networks of physically interacting proteins, to the best of
our knowledge the first direct evidence that evolution drives the modularity
of PPI networks was provided by Wuchty et al. (2003). They looked beyond
a single protein pair and studied the more complex patterns of interacting
2.2. Unravelling the relations between evolution and PPI network structure

proteins, called topological motifs. In general, they found that, as the number of nodes in a motif and number of links among its constituents increase, a greater and stronger conservation of the proteins could be observed. This was corroborated by Vergassola et al. (2005) who focused on specific instances of motifs known as cliques. Cliques are topological patterns where all protein constituents interact with each other. Vergassola et al. (2005) provided evidence for co-operative co-evolution within cliques of interacting proteins. Later, Lee et al. (2006) investigated motifs at a higher resolution level, by defining for each motif different motif modes based on functional attributes of interacting proteins: again their findings indicated that motifs modes may very well represent the evolutionary conserved topological units of PPI networks. More recently, Liu et al. (2011) studied network motifs according to the age of their proteins and discovered that the proteins within motifs whose constituents are of the same age class tend to be densely interconnected, to co-evolve and to share the same biological functions. Moreover, these motifs tend to be within protein complexes.

The finding that modularity of PPI networks is constrained by evolution and that conserved interactions are enriched in dense motifs and regions of a PPI network also suggest that protein complexes present in such cohesive areas should be evolutionary driven (Jancura et al., 2012). As putative protein complexes can be extracted from a PPI network by means of clustering techniques, Jancura et al. (2012) detected such protein complexes in the PPI network consisting of only yeast proteins having an ortholog in another organism and compared them with those protein complexes derived either by using the global topology of a yeast PPI network or by using a network induced by randomly selected proteins. The in-depth examination of enriched functions in these three types of protein complexes revealed that evolutionary-driven complexes are functionally well differentiated from other two types of protein complexes found in the same interaction data. As a consequence, new complexes and protein function predictions could be unravelled from PPI data by using a standard clustering approach with the inclusion of evolutionary information. In addition, evolutionary-driven complexes were found to be differentially conserved, in particular some complexes were detected for all distinct set of orthologs.
as determined by comparison with different species, some exhibited only a subset of proteins identifiable in a complex across all species, and some complexes being found only for one specific set of orthologs. This suggests that presence of evolution in modularity of PPI networks is more versatile and flexible with different degrees of conservation.

The findings of Jancura et al. (2012) seem to conform with related studies that focused on evolutionary cohesiveness of protein functional modules in order to investigate whether a group of proteins which functionally interact, co-evolve more cohesively than a random group of proteins. Either known protein complexes and pathways were analysed (Fokkens and Snel, 2009, Seidl and Schultz, 2009, Snel and Huynen, 2004) or putative protein modules usually derived from integrated networks of functional link evidences (Campillos et al., 2006, Zhao et al., 2007, Zinman et al., 2011). A different strategy was employed by Yamada et al. (2006) who at first detected evolutionary modules which were afterwards compared with enzyme connectivity in a metabolic network.

Although the co-evolution of modules is assessed by the presence or absence of modules’ constituents across a set of species, there is no standard method to measure the degree to which a module evolves cohesively (Fokkens and Snel, 2009). For instance, Snel and Huynen (2004) used the deviation of the number of modules’ orthologs per species from the average number of modules’ orthologs per species, whereas Campillos et al. (2006) measured the fraction of joined evolutionary events given the reconstructed, most parsimonious evolutionary scenario of the genes in a module over their phylogenetic profiles.

Despite this measures’ diversity, the common conclusion is that the majority of modules evolve flexibly (Campillos et al., 2006, Fokkens and Snel, 2009, Seidl and Schultz, 2009, Snel and Huynen, 2004, Yamada et al., 2006). Also, it appears that curated modules evolve more cohesively than modules derived from high throughput interaction data (Fokkens and Snel, 2009, Seidl and Schultz, 2009, Snel and Huynen, 2004). Moreover, there is a different enrichment in functions which co-evolve. For example, biochemical pathways, certain metabolic and signalling processes, as well as core functions like transcription and translation, tend to have higher rate of
2.2. Unravelling the relations between evolution and PPI network structure

evolutionary cohesiveness (Campillos et al., 2006, Fokkens and Snel, 2009, Zhao et al., 2007). This is also supported by methods which cluster phylogenetic profiles in order to detect biochemical pathways or to predict functional links and thus exploiting the predictive power of phylogenetic methods (Glazko and Mushegian, 2004, Li et al., 2009, Watanabe et al., 2008). These methods show a relatively good performance in characterizing biochemical pathways but seem to have a limited coverage for physically interacting proteins (Watanabe et al., 2008). A dubious result was reported on inter-connectivity of cohesive and flexible modules. Specifically, Fokkens and Snel (2009) demonstrated that components of cohesive modules are less likely to interact with each other than in the case of flexible modules, while two other studies (Campillos et al., 2006, Zinman et al., 2011) suggest cohesive modules to be more highly connected.

It is possible that the above studies underestimated the actual degree of evolutionary cohesiveness present in the modularity of protein interaction networks due to their conservative approach, the limitations in ortholog detection as well as the cohesiveness measures which are restricted to phylogenetic profiles. Nevertheless, they show that, as evolution is a complex process, its presence in modularity of protein interaction networks also exhibits a very complex nature, whose understanding is far from being complete. Evolution itself, indeed, can be expected to be asynchronous and heterotactous along the tree of life.

In general, the interim evidence shows different evolutionary pressure for different types of protein interactions and data. In particular, the slowly evolving interacting partners are enriched in stable, permanent complexes, and functional modules such as biochemical pathways and curated complexes exhibit higher evolutionary cohesiveness than high throughput complexes. It seems that the co-evolutionary degree of modules within PPI networks increases with greater integration of various sources of evidence for proteins to functionally interact (Zinman et al., 2011). Also, not all protein complexes and functional modules need to be co-evolutionary modules (Fokkens and Snel, 2009). There is a continuum from extremely conserved to rapidly changing modules, where those modules found to be co-evolving appear to be enriched in certain, specific functional categories (Campillos
et al., 2006). In addition, the degree of conservation and co-evolution of functional modules within interaction networks seem to reflect cellular organization and their spatio-temporal characteristics. For instance, cohesive modules can be classified according to their evolutionary age as ancestral, intermediate and young, where one may observe ancient, ancestral modules to be highly conserved and perform essential, core processes such as information storage and metabolism of amino acids, while young modules are less conserved and responsible for the communication with the environment (Campillos et al., 2006). Therefore one might expect ancestral modules to contain static, obligate interactions as the proteins of essential functions tend to involve multiple domains with slow evolutionary rates, whereas young modules can be enriched with dynamic, transient interactions with less but fast evolving protein domains to allow adaptation to the environment.

2.3 Using evolutionary information for knowledge discovery in PPI networks

The tendency of functionally linked or physically interacting proteins and densely interacting motifs to exhibit correlated evolution and/or to be conserved across species is at the core of methods for inferring relevant biological information using PPI networks. Although such biological information can be limited and biased towards specific type of known interactions and protein functions, it allows one to infer new, unknown functions of proteins, to improve the understanding of biological systems, and to guide the discovery of drug-target interaction. In its basic form, the knowledge discovery process is based on the transfer of information involving a single interaction between two organisms, while in its most complex form it involves the identification and transfer of protein complexes across multiple species. In the sequel we summarize concepts and techniques used to achieve these goals, in particular the notions of “interologs” and of multiple PPI networks alignment.
2.3. Predicting protein interaction: Interologs

If two proteins physically interact in one species and they have orthologous counterparts in another species, it is likely that their orthologs interact in that species too. If such conserved interactions exist, they are called interologs. This simple method of protein interaction inference was firstly introduced and tested by Walhout et al. (2000) on proteins involved in vulval development of nematode worm, where potential interactions between these proteins were identified based on interactions of their orthologs in other species. Later, Matthews et al. (2001) performed a large-scale analysis of this inference technique using the yeast PPI network as a model and proteins of worm as a target. Although the success rate of detection of inferred interactions by Y2H analysis was between 16%-31%, it represented a 600-1100-fold increase compared to a conventional approach at that time (Matthews et al., 2001).

The interologs-based protein interaction prediction has become one of the standard methods for in silico PPI prediction. The method can be easily extended to more PPI data from multiple species. In particular, having two groups of orthologs, where each ortholog group contains proteins from the same N species, and observing an interaction between proteins of these orthologous groups in (N−1) species, the interaction between proteins of the N-th species present in the ortholog groups can be predicted (see Figure 2.2). This multidimensional character of interolog inference has been extensively used to predict and build databases of the whole interactome for various species, either as a stand alone approach or in combination with other in silico methods, which often integrate multiple data types including the gene co-expression, co-localization, functional category, the occurrence of orthologs and other genomic context methods (Brown and Jurisica, 2005, Geisler-Lee et al., 2007, Gu et al., 2011, He et al., 2008, Huang et al., 2004, Jonsson and Bates, 2006, Lehner and Fraser, 2004, Pavithra et al., 2007, Persico et al., 2005, Titz et al., 2008, Yellaboina et al., 2008). In this way researchers could provide the first sketch of human interactome (Lehner and Fraser, 2004), build the interactome of plants (Geisler-Lee et al., 2007, Gu et al., 2011), and improve the understanding of processes in a malarial
Figure 2.2: The interolog transfer across $N$ species ($S_1, S_2, \ldots, S_N$). Dark grey nodes are proteins. Solid edges are interactions between proteins of orthologous groups $A = \{A_1, A_2, \ldots, A_N\}$ and $B = \{B_1, B_2, \ldots, B_N\}$. Dash edges link the proteins of the same orthologous group across species. The dot edge between proteins $A_N$ and $B_N$ indicates the interolog-based prediction of the interaction between the proteins in the $S_N$ species.

parasite (Pavithra et al., 2007) or in cancer (Jonsson and Bates, 2006). Also, three, up-to-date, tools have been recently implemented and made available to perform this inference task (Gallone et al., 2011, Michaut et al., 2008, Pedamallu and Posfai, 2010).

Several algorithmic enhancements of the interologs-based approach have been introduced since the first proposal of a systematic use of interolog inference (Matthews et al., 2001). For instance, Yu et al. (2004b) have strengthened the definition of ortholog by using a reciprocal best-hit approach and compared it to the original one-way best-hit approach implemented by Matthews et al. (2001). In addition, they required a minimum level for a joint similarity of orthologous sequences in order to perform interolog mapping. Their method yielded a 54% accuracy in contrast to a 30% of the previous method by Matthews et al. (2001).

Other approaches exploited the knowledge on a higher conservation rate of PPIs in dense network motifs. For instance Huang et al. (2007) scored interologs according to the density of the topological pattern containing
the respective PPI of the interolog in a model species as determined by
the extraction of maximal quasi-cliques from the PPI network of the model
species. This score was integrated with scores of other various features
used for PPI prediction, such as tissue specificity, sub-cellular localization,
interacting domains and cell-cycle stage. The use of multiple types of fea-
tures was shown to yield more accurate predictions of PPIs in comparison
with other interolog-based methods used to build interactome databases.
More recently, Jaeger et al. (2010) proposed another interesting method
based on two steps. First a set of all candidate interologs is built across the
considered species. Next, interologs are assembled into maximal conserved
and connected patterns by detecting frequent sub-graphs appearing in the
interolog network of the candidate set. Only functionally coherent patterns
were used for interolog inference.

The interolog concept was also modified and used in other ways and
application domains. In particular, Tirosch and Barkai (2005) proposed a
method to assess and increase the confidence of a predicted PPI by examin-
ing the co-expression of proteins of its potential interolog in other species.
Chen et al. (2007) extended interolog mapping for homologous inference of
interacting 3D-domains and they built a database of so-called 3D-interologs
(Lo et al., 2010a). Chen et al. (2009) used interologs to transfer conserved
domain-domain interactions. Recently, Lo et al. (2010b) combined this in-
terolog domain transfer with the former 3D-interolog detection technique
and implemented an integrated tool for searching homologous protein com-
plexes. Finally, Lee et al. (2008) exploited interologs to predict inter-species
interactions.

Despite the successful use of interolog inference, a gap was observed
between the actual, observed number of conserved interactions and the ex-
pected theoretical coverage (Gandhi et al., 2006, Lee et al., 2008). In order
to test the reliability of interolog transfer, Mika and Rost (2006) performed
a comprehensive validation of the method on several datasets. Their find-
ings suggested that interolog transfers are only accurate at very high levels
of sequence identity. In addition, they also compared the interolog transfer
within species and across species. In the case of within-species interolog
inference a PPI is transferred onto proteins which are sequence similar to
the proteins of the considered PPI in the same species. Surprisingly, such paralogous interolog transfers of protein-protein interactions were shown to be significantly more reliable than the orthologous ones. This result was later substantiated by Saeed and Deane (2008), indicating that homology-based interaction prediction methods may yield better results when within-species interolog inference is also considered. In addition, Brown and Jurisica (2007) argued that one also needs to take into account whether all interactions have equal probability of being transferred between organisms. For example, the dynamic components of the interactomes are less likely to be accurately mapped from distantly related organisms. Moreover, there is apparent bias of interologs to be enriched in stable, permanent complexes (Brown and Jurisica, 2007), which is completely in accordance with findings on the different evolution of transient and permanent interactions. On the other hand, it is likely that the performance of interolog inference be under-estimated since its accuracy is assessed using experimentally-verified interactions based on Y2H techniques or high-throughput datasets with a high abundance in Y2H interactions, which were found to be highly enriched in transient and inter-complex connections (Yu et al., 2008).

### 2.3.2 Pairwise protein network alignment

Detection and transfer of an interolog between species have motivated the study and exploration of interspecies conservation of protein interactions on a global scale. In particular, instead of focusing on a conserved interaction alone one can compare and align whole interactome maps of distinct species, which mimics the idea behind sequence alignment methods. This approach gave a rise to so-called *network alignment* problem (Sharan and Ideker, 2006).

Using protein network alignment, one can either search for conserved functional network structures such as protein complexes and pathways, or identify functional orthologs across species. As a result this approach should provide a greater evidence and support for protein function and protein interaction prediction for yet uncharacterized or unknown biological processes. Protein network alignment methods can be classified into two main
2.3. Using evolutionary information for knowledge discovery in PPI networks

groups: local network alignments and global network alignments.

As most of the research attention has focused on comparing PPI networks of two different species, here we discuss the successive development of methods for, so-called, pairwise network alignment. In sequel we survey local pairwise alignments for detecting evolutionary conserved pathways, local pairwise alignments for detecting conserved protein complexes, and global pairwise network alignment techniques.

Local pairwise network alignment for pathway detection and query tasks

The main goal of local protein network alignment is to detect conserved pathways and protein complexes across species, by searching for local regions of input networks having both high topological similarity between the regions and high sequence similarity between proteins of these regions. The standard approach to this task consists of two main phases: an alignment phase and a searching phase. In the first phase a merged network representation of compared PPI networks is constructed, called alignment or orthology graph. The second phase performs a search for the structures of interest in the orthology graph. Each output result corresponds to a pair or multiplet of complexes or pathways which are evolutionary conserved across the two or more (PPI networks of the) species, respectively (see Figure 2.3).

The first alignment method of whole PPI networks of two species using protein sequence similarity was introduced by Kelley et al. (2003). In this method, called PathBLAST, first a many-to-many mapping between proteins of the two species is determined by considering each pair of proteins with a sequence similarity higher than a given threshold as putative orthologs. Next, every orthologous pair is encoded in one alignment node of the new alignment graph and three types of edges (direct, gap and mismatch edge) are identified between these alignment nodes as follows. The direct edge corresponds to the case when a PPI between proteins of two orthologous pairs exists in the PPI networks of both species. The gap edge represents the case when in one species the respective proteins of alignment’s nodes are connected indirectly through a common neighbour.
Figure 2.3: A hypothetical result of a pairwise local protein network alignment between species $S_1$ and $S_2$. Solid edges are protein-protein interactions within one species. Dash edges are putative orthologous mapping between proteins of the two species. Dark nodes on the grey background and dark edges of both types indicate the solution of local network alignment, that is the evolutionary conserved complexes and respective orthologous mapping. Grey nodes and edges represent the rest of the networks and orthologous relationships.
Finally, the mismatch edge between alignments nodes is formed if such indirect connection is found between the corresponding proteins in the PPI networks of both species. Gap and mismatch edges are used to describe possible evolutionary variations or account for experimental errors in data (Kelley et al., 2003). In the search phase, the alignment graph is turned into acyclic sub-graphs by random removal of alignment edges, which allows to extract high-scoring paths in linear time by a dynamic programming approach. The score of a path is computed as the sum of log probabilities of true orthology encoded in alignment nodes of the path and of true conserved interactions encoded by alignment edges contained in the path. Interestingly, the method was also applied to align a PPI network with its own copy. In this way they could identify conserved (paralogous) pathways within one species.

The work of Kelley et al. (2003) was followed by other alignment techniques for discovering conserved pathways based on evolutionary conservation. The main drawbacks of PathBLAST are that it detects conserved linear pathways in protein interaction data, which is represented as an undirected graph, and it has an exponentially worsening efficiency with the expected increasing length of a pathway to be detected. To circumvent these limitations Pinter et al. (2005) proposed an alignment technique designed explicitly for metabolic networks with directed links between enzymes. The method also handles more complex structures than a simple path, because the scoring of the alignment is based on sub-tree homeomorphism, which can be solved by an efficient deterministic approximation. Another enhancement for the pathway alignment problem was proposed by Wernicke and Rasche (2007) who designed a method that does not impose topological restrictions upon pathways and exploits the biological and local properties of pathways within the network. Another effective approach to metabolic network alignment was developed by Li et al. (2008) which uses an integrative score on compound and enzyme similarities. Pathway alignment has been further extensively investigated and various other techniques have been proposed (see e.g. Cheng et al., 2008, Koyutürk et al., 2006a, Li et al., 2007).

The evolutionary mapping of PathBLAST can also be used to query
a known pathway of one species into the PPI network of another species. However, due to limitations and algorithmic constraints of PathBLAST, many other methods have been developed with a focussed application of orthologous querying of biological functional complexes, and tools and web-services are available for querying general pathways and other types of protein functional modules across species (see e.g. Blin et al., 2009, Bruckner et al., 2009, Dost et al., 2008, Qian et al., 2009, Shlomi et al., 2006, Yang and Sze, 2007).

**Local pairwise network alignment for protein complex detection**

Another group of methods which followed PathBLAST focus on detection of conserved protein complexes across (PPI networks of two or more) species. As these methods compare networks of physical interactions, the identified complexes can be used for interolog prediction as well as for protein function prediction of yet uncharacterized proteins. The detected conserved complexes are either (putative) entire physical complexes or conserved parts of them (Figure 2.3).

To the best of our knowledge, the first method for detecting conserved complexes using pairwise comparison of PPI networks was introduced by Sharan et al. (2004, 2005a) and called NetworkBLAST. It can be viewed as a direct extension of PathBLAST for the task of complex detection across species. The method employs a comprehensive probabilistic model for conservation of protein complexes and searches for heavy induced subgraphs in the weighted orthology graph. As the maximal induced sub-graph problem is computationally intractable, NetworkBLAST employs a bottom-up greedy heuristic for this task.

Many alignment network techniques which followed NetworkBLAST are motivated by the computational intractability issue derived from the problem of a finding maximal common or induced sub-graph in an ortholog graph, and are based on different heuristics. For instance, Koyutürk et al. (2006b) partitions the alignment graph into smaller clusters by performing an approximated balanced ratio-cut. In another method by Koyutürk et al. (2006a) the most frequent interaction motifs are extracted from an
orthology-contracted graph. Liang et al. (2006) transforms the problem of maximal common sub-graph into the problem of finding all maximal cliques in the graph. Recently, Tian and Samatova (2009) introduced an algorithm based on detection of connected-components of the orthology graph solvable in a very efficient way.

Other researchers propose to restrict the search space to cope with intractability issue of searching phase instead of performing heavy heuristics. For example Li et al. (2007) pre-clusters one PPI network in order to detect candidate complexes which are afterwards aligned to the target species network with an exact integer programming algorithm. Janura and Marchiori (2010) proposed a pre-processing algorithm based on detection of network hubs for dividing PPI networks, prior to their alignment, into smaller sub-networks containing potential conserved modules. Each possible pair of sub-networks can be later aligned with a state-of-the-art alignment method where the search phase can be performed by means of an exact algorithm, allowing one to perform network comparison in a fully modular fashion and possibly to parallelize the computation. An interesting modular approach was introduced by Narayanan and Karp (2007), where an orthology graph is not constructed but rather networks are compared and split consecutively in several recursive steps until all possible solutions, conserved sub-graphs, are found. Similarly, Gerke et al. (2007) only compares, but does not merge, local hub-centred regions of PPI networks as identified by clustering coefficients and node degrees. The method by Ali and Deane (2009) is again another example of approach where an alignment graph is not explicitly constructed; there interspecies protein similarities are considered as new edges in such a way that species PPI networks and similarity edges between them are encoded into a single global meta-graph which can be searched by standard clustering techniques.

There are also alignment methods which try to incorporate or use other types of information than just the one based on sequence similarity and interaction conservation. For instance, Guo and Hartemink (2009) exploited the findings on co-evolving interacting domains which mediate PPI’s and, instead of using putatively homologous proteins for alignment, compares PPI networks across species according to conserved domains of
protein-protein interactions. Ali and Deane (2009) propose a functionally guided alignment of PPI networks, where a scoring function incorporates not only sequence and topological similarity of aligned proteins but also their gene co-expression characteristics and coherence of functional annotations. Thus, the method can be seen as detecting functional modules shared across species rather than strictly evolutionary modules. Finally, Berg and Lässig (2006) developed a generalized Bayesian alignment method applicable to different biological networks.

Despite various pairwise alignment techniques have been introduced, only a few of them embody an evolutionary model of PPI networks in the scoring scheme of an alignment. Notably, Koyutürk et al. (2006b) were the first to introduce a method that builds the orthology graph following the duplication/divergence model based on gene duplications. Another interesting method was proposed by Hirsh and Sharan (2007) who extended the probabilistic score of NetworkBLAST to assess the likelihood that two complexes originated from an ancestral complex in the common ancestor of the two species being compared under the evolutionary pressure of duplication and link dynamics events.

**Global pairwise network alignment**

In contrast to local network alignment, which uses many-to-many homologous mapping between proteins of distinct species to detect local conserved regions of a high topological similarity in the respective PPI networks, global protein network alignment uses this mapping to define an unique, globally optimal mapping across whole topologies of PPI networks (Singh et al., 2007), even if it were locally suboptimal in some regions of the networks. In the most strict form of this unique mapping each node in one input network is either matched to one node in the other input network or has no match in the other network (see Figure 2.4). Thus the goal of global protein network alignment is to define functional orthologs across species, as the solution offers a way to resolve the ambiguity of orthology detection with the use of species interactome map. Naturally, as a by-product the global alignment can also identify conserved complexes or pathways.
2.3. Using evolutionary information for knowledge discovery in PPI networks

Figure 2.4: A hypothetical result of a global pairwise protein network alignment between species $S_1$ and $S_2$. Solid edges are protein-protein interactions within one species. Dash edges are putative orthologous mapping between proteins of the two species. Dark nodes and dark edges of both types indicate the solution of global network alignment, that is one-to-one orthologous mapping (dark dash edges) between proteins corresponding to functional orthologs (dark nodes) and conserved protein-protein interactions (dark solid edges) between those proteins in respective species. Grey nodes and edges represent the rest of the networks and orthologous relationships.
Chapter 2. A survey on evolutionary analysis in PPI networks

To the best of our knowledge, the first method performing explicitly global alignment on pair of networks, called IsoRank, was introduced by Singh et al. (2007). Similarly to the local network alignment problem, the global network alignment problem is in general computationally intractable. As a consequence, IsoRank employs an approximation using an eigenvalue framework in a manner analogous to Google’s PageRank algorithm.

Several advancements have naturally followed the introduction of IsoRank. For instance, Evans et al. (2008) proposed an asymmetric network matching algorithm based on a network simulation method called quantitative simulation, where a similarity score of a protein pair is iteratively updated by the similarity scores of their neighbours and vice versa until a unique global optimum is found. Other researchers focused more on formulating global alignment as combinatorial optimization problems. For instance Zaslavskiy et al. (2009) redefined the problem of global alignment as a standard graph matching problem and investigated methods using ideas and approaches from state-of-the-art graph matching techniques. Klau (2009) formalized global network alignment as an integer linear programming problem, where a near-optimal solution with a quality guarantee is found by solving a Lagrangian relaxation of the original optimization formulation. Recently, Chindelevitch et al. (2010) proposed a method where the global alignment is encoded as bipartite matching and applied a very efficient local optimization heuristic used for the well-known Travelling Salesman Problem.

2.3.3 Multiple protein network alignment

The methods on network alignment discussed so far perform alignment of two PPI networks of distinct species. The next natural extension is aligning more than two PPI networks, that is multiple network alignment. A first attempt to perform multiple local network alignment using three species was done by Sharan et al. (2005b), which exploited the scoring model of NetworkBLAST. However, the method scales exponentially with the number of input species and consequently it is ineffective for large scale comparisons.
Apart from the scalability problem, there are also other issues related to the problem of aligning more than two species. For instance, the putative orthologous mapping of certain proteins does not need to span across all species, meaning that proteins may be conserved only for a particular subset of species. This “orthology decay” is more evident when a large number of increasingly distant species are considered in the alignment. As a result, functional modules, such as pathways and complexes, can have a different degree of conservation, with some modules being strictly conserved across all species and some other modules being conserved only for a particular clade. Thus, a good alignment method should allow one to search for conserved modules at different degrees of conservation. However, such requirement also increases the complexity of searching and consequently one may need to prune the number of all possible species combinations in alignment.

To the best of our knowledge, the first method capable of an efficient comparison of multiple PPI networks, called Graemlin, was introduced by Flannick et al. (2006). The alignment model of the method allows one to perform local as well as global alignment and is also applicable for querying tasks of particular biological modules of interest across PPI networks. It employs a rather involved scoring scheme which allows one to search for conserved pathways as well as for conserved complexes. It also outputs modules with a different conservation degree. Graemlin progressively aligns the closest pair of PPI networks according the species distance measured using a phylogenetic tree, until the last pair on the root of the tree is compared, corresponding to the most conserved parts of the aligned networks. The main disadvantage of this approach is that it involves to estimate many parameters. Recently, a supervised, automated parameter learner was proposed to lessen the burden of parameter tuning (Flannick et al., 2009).

Another phylogeny-guided local network alignment was proposed by Kalaev et al. (2008). Although the method uses the same probabilistic scoring for conserved complex as NetworkBLAST, it avoids its exponential scalability by redefining the alignment model such that it does not construct the merged representation of aligned networks but represents them as separate layers interconnected via orthologous mapping. Then a seed, that
is, a group of putatively orthologous proteins spanning across all species, is selected using the species phylogeny and greedily expanded by adding other proteins being orthologous to each other in all respective species in order to maximize the alignment conservation score. The proposed method, however, identifies only protein complexes conserved across all species and does not detect complexes conserved only for a certain subset of species.

Notably, the functionally guided network alignment method Ali and Deane (2009), previously mentioned as one of the methods for pairwise alignment, was also shown to perform efficiently local alignment of multiple networks.

All these multiple local network alignments do not reconstruct a plausible evolutionary history of PPI networks based on a model of evolution, although they might be phylogeny-aware. Motivated by this observation, Dutkowski and Tiuryn (2007) introduced a new multiple local network alignment method, called \textit{CAPPI}, which from the given PPI networks of distinct species aims to reconstruct an ancient PPI network of the common ancestor. The method uses a Bayesian inference framework based on a duplication and divergence model of network evolution which mimics the processes by which most protein interactions are formed. After the reconstruction step, the ancestral network is decomposed into connected components which correspond to the ancestral modules of protein interactions and are projected back to the original networks to obtain the actual conserved network residues. Although the demonstrated application of the method was restricted to orthologous groups spanning across all species (Dutkowski and Tiuryn, 2007), to the best of our knowledge \textit{CAPPI} is the only model-based approach for large-scale ancestral network reconstruction.

Among the multiple alignment methods above mentioned, only \textit{Graemlin} was shown to perform a global multiple network alignment, yet it relies on a involved parameter estimation step and phylogeny-guided approximation. Recently Liao et al. (2009) developed another global alignment technique which is fully unsupervised and phylogeny-free. The method, called \textit{IsoRankN}, is built on the \textit{IsoRank} algorithm mentioned above (Singh et al., 2007) and its extension to the multiple global network alignment (Singh et al., 2008a). At first \textit{IsoRankN} scores topological and sequence similar-
2.3. Using evolutionary information for knowledge discovery in PPI networks

ity matching between putatively orthologous proteins of each pair of input networks using IsoRank. Then, a maximum k-partite graph matching problem is formulated on the induced graph of pairwise alignment scores (Singh et al., 2008a) and the exact solution is approximated by a spectral graph partitioning algorithm. IsoRankN also effectively identifies one-to-one orthologous mappings for all subset of species and appears to out-perform Graemlin in terms of coverage and quality of functional enrichments.

2.3.4 Applications and future developments

Local and global alignment methods have been successfully applied to study evolution of species and to discover relevant biological knowledge. For example, Suthram et al. (2005) applied the network alignment of Sharan et al. (2005b) to examine the degree of conservation between the Plasmodium protein network and other model organisms, such as yeast, nematode worm, fruit fly and the bacterial pathogen Helicobacter pylori. They investigated whether the divergence of Plasmodium at the sequence level is reflected in the configuration of its protein network. Indeed, the alignments showed very little conservation suggesting that the patterns of protein interaction in Plasmodium, like its genome sequence, set it apart from other species (Suthram et al., 2005).

Another application of local network alignment was performed by Tan et al. (2007) who combined transcriptional regulatory interactions with protein-protein interactions and identified co-regulated complexes between yeast and fly revealing different conservation of their regulators. This finding advocates that PPI networks may evolve more slowly than transcriptional interaction networks. In addition, Schwartz et al. (2009) and Dutkowski and Tiuryn (2009) used conserved complexes detected by network alignments for protein interaction prediction in a manner similar to the interologs transfer approach and demonstrated their usefulness. In particular, Schwartz et al. (2009) provided a comprehensive experimental design which includes PPI prediction using network alignment, and demonstrated how effectively it reduces the cost of interactome mapping.

Furthermore, Bandyopadhyay et al. (2006) presented the first system-
atic identification of functional orthologs based on protein network comparison. They used the pairwise local alignment model of Kelley et al. (2003) to construct the orthology graph and then they resolved ambiguity of orthology mapping by fitting a logistic function previously trained on a known set of functional orthologs. In contrast, Singh et al. (2008b) predicted functional orthologs in unsupervised manner by using explicitly a global multiple network alignment method.

Finally, Kolar et al. (2008) performed a cross-species analysis of two herpes-viruses using the generalized Bayesian network alignment of Berg and Lässig (2006). Interestingly, the performed alignment employs in its probabilistic scoring system evolutionary rates of sequences and thus it goes beyond the narrow use of orthologous mapping as done in all other alignment techniques. The method predicted meaningful functional associations that could not be obtained from sequence or interaction data alone.

Despite the recent progress and increasing number of network alignment tools, their further development remains an ongoing research issue, in particular for multiple network comparison. Only a few methods perform the scoring of alignment according to evolutionary models and there is only one of them which fully reconstructs network evolutionary history. This clearly is in contrast with the numerous techniques for the reconstruction of evolutionary history of gene families. Also, actual alignment methods do not distinguish among diverse types of interactions, specifically between transient and permanent interactions. For example, the prior knowledge on different evolutionary behaviour of these types of physical interactions could be incorporated into a scoring scheme of alignment construction.

In addition, all but one network comparison methods just rely on the straightforward use of putative orthologous mapping as identified by sequence comparison or available in orthologous databases, but they do not employ evolutionary measures, such as evolutionary distances or retentions, which can be derived from the corresponding sequence alignments. These measures assess the level of evolutionary conservation and they could potentially improve the performance of network alignments.

Mostly all current applications of network alignments have worked with networks of physical interactome. However, the power of network alignment
2.3. Using evolutionary information for knowledge discovery in PPI networks

for functional annotation and other system biology applications could be explored when one performs comparison of more general, functional interaction networks. One may expect that such alignment could reveal a higher number of conserved modules as the interspecies conservation of modularity across protein networks increases with combined, integrated evidence for a pair of proteins to be functionally linked. Finally, all available methods here considered focused on conservation of modules but not on the more general concept of module evolutionary cohesiveness or co-evolution. The evolutionary cohesiveness can be assessed especially for the case of multiple alignments. Indeed, all conserved modules are inherently very cohesive, however not all evolutionary modules need exhibit the correlated conservation at a level as expected by actual multiple network alignments. Protein functional modules differ in the degree of conservation and also in the degree of cohesiveness.
Chapter 3

Quantifying the orthology signal in a PPI network at a protein complex level

Stable evolutionary signal has been observed in a Yeast PPI network suggesting more connected regions of a PPI network to be potential mediators of evolutionary information. Because such regions of PPI networks contain functional complexes, we are motivated to exploit the orthology relation for identifying complexes that can be clearly attributed to such evolutionary signal and we propose a computational methodology for detecting the orthology signal present in a PPI network at a functional complex level. Its application to the Yeast PPI network indicated that evolutionary information can be retrieved from the structure of the network at this level.\textsuperscript{1}

\textsuperscript{1}This chapter is based on the following works:


Chapter 3. Quantifying the orthology signal in a PPI network

3.1 Introduction

Analysing and mining protein-protein interaction (PPI) networks data using evolutionary information is a central research area in bioinformatics (see e.g. Brown and Jurisica, 2007, Campillos et al., 2006, Erten et al., 2009, Fokkens and Snel, 2009, Sharan and Ideker, 2006, Vespignani, 2003, Woźniak et al., 2010, Wuchty et al., 2003, 2006, Yosef et al., 2009). In this context, evolutionary information is usually described by means of the orthology relation between proteins. In general, two proteins are orthologous if they originated from a common ancestor, having been separated in evolutionary time only by a speciation event. Orthologous proteins have high amino acid sequence similarity and usually retain the same or very similar function, which allows one to infer biological information between the proteins. Obviously, orthology as such is very important in studying evolution. Therefore, the problem of establishing proper orthology relations has been widely studied in comparative genomics (see for instance Kuzniar et al., 2008) and many databases and public resources of orthologs have been made available, such as Inparanoid (Berglund et al., 2008, O’Brien et al., 2005) and OrthoMCL-DB (Chen et al., 2005).

A recent study performed by Wuchty et al. (2006) used such available orthology information for detecting stable evolutionary signal in a yeast PPI network. This signal was extracted at a protein-protein interaction level, using pairwise orthologs with respect to various different species. The authors observed that a high local clustering around protein-protein interactions correlates with evolutionary conservation of the participating proteins. This means that highly connected proteins and protein pairs embedded in a well clustered neighbourhood tend to be evolutionary conserved and therefore retain their evolutionary signal. These findings suggest also that more connected regions of a PPI network are potential mediators of evolutionary information.

Motivated by the above observations, in this study we focus on the explicit use of orthology for detecting evolutionary signal at a functional complex level, that is, functional complexes that can be clearly attributed to this evolutionary signal. To this aim, we try to characterize functions...
of those complexes predicted by clustering the sub-graph of a PPI network induced by all proteins having orthologs in another given species, but not predicted (or predicted for a smaller fraction of proteins) neither when clustering the entire network, nor when clustering sub-graphs of the network induced by random sampling of proteins. We consider the resulting functions as a strong characterization of the underlying evolutionary signal of orthologs at functional complex level, since they are suppressed or not observed when clustering using the entire network and are not outcomes of a stochastic process.

Specifically, given the yeast PPI network and proteins from another species, we apply a state-of-the-art clustering algorithm to (1) the yeast PPI network, (2) the sub-network of the yeast PPI network induced by selecting only proteins with ortholog in the considered other organism, and (3) the sub-networks of the yeast PPI network induced by sampling a given number of proteins at random. In this way we generate three classes of clusters called $GC$ (global clusters), $OC$ (ortholog clusters) and $RC$ (random clusters). Note that the latter class of clusters is the collection of cluster sets produced by the application of clustering to the PPI network induced by a random selection of a set of proteins (of size equal to that of the set of proteins used to generate the $OC$ class) repeated for dozens times. For all clusters in each class we infer putative functions by measuring their gene ontology (GO) functional enrichment (Ashburner et al., 2000) using only experimentally validated annotations, and consider as putative protein complexes only those clusters with a putative function that is significantly coherent within the corresponding cluster.

The putative complexes of the $GC$ class represent results globally observable in the whole interaction data without any additional information and hence play also a suppressor of any potential external biological signal present in the data. The putative complexes of the $RC$ class simulate a random signal of the given protein sample size in the protein interaction data. Thus, the $OC$ class complexes may be attributed to the orthology signal only when their functionality clearly differentiates from those of $GC$ and $RC$ class.

To this end, for a set of complexes and a certain function, we compute
the functional retrieval index as the fraction of proteins contained in the complexes and having the function experimentally validated with respect to the set of candidate proteins having the function also experimentally validated and from which complexes were derived. This fraction quantifies the presence of that function in a given protein complex set. This allows us to identify functions whose proteins’ fraction is higher in complexes from the \( OC \) class than in complexes from the other two classes. Consequently, we consider the corresponding complexes in \( OC \) class as describing the orthology signal (with respect to the considered species). Furthermore, we analyse those complexes in the \( OC \) class having a predicted function for its proteins that is not inferred when using complexes of \( GC \) class. Finally we discuss the new meaningful functions for well-defined as well as for unknown proteins that are present in the compilation of putative complexes.

In previous works on phylogenetic analysis of protein networks and complexes evolutionary information was usually used as a mean for evaluating the preservation of orthology information in functional modules (Campillos et al., 2006, Erten et al., 2009, Fokkens and Snel, 2009, Wuchty et al., 2003, Yosef et al., 2009). Here, however, we incorporate evolutionary information beforehand and perform a comparative differential analysis for detecting evolutionary signal at complex, functional level. Our identification of protein complexes uses only the topology of the network of the considered species and orthology information from another species, without requiring knowledge on the interactome of the other species.

In general, our approach differs from comparative network methods (Sharan and Ideker, 2006), as the latter aim to find evolutionary conserved modules across species, thus exploiting both orthology and network topology of the considered organisms. The clusters we obtain are in one species and are related to the orthology signal with respect to another species, but are not required to be evolutionary conserved through species (we do not enforce any type of similarity at the graph-structure level). Furthermore, comparative methods mostly do not use ‘known’ orthologs in available databases but rather they rely on sequence similar proteins, where the level of required similarity is determined by a minimal similarity score threshold. Instead, our method exploits the orthology information avail-
3.2 Methods

3.2.1 Data

We performed the analysis on a widely used and well-studied species, namely *Saccharomyces cerevisiae* (yeast), since its PPI network is one of the best characterized and the functionality of its proteins has been extensively studied. This makes yeast a good standard model species for protein network analysis.

Specifically, we used the budding yeast interaction data collected by Georgii et al. (2009). The data combines interaction data from DIP (Xenarios et al., 2002) and MPact (Guldener et al., 2006), and interactions from the core datasets of the TAP mass spectrometry experiments (Gavin et al., 2002, Krogan et al., 2006). This yeast interaction data are weighted by the method proposed by Jansen et al. (2003) to measure the confidence of interactome. As a result, the low confidence interactions are ignored and

able in existing databases. Moreover, the study we propose differs from those aiming to find evolutionary conserved modules across species because their output results in cluster multiplets derived from all PPI networks of multiple species being compared, where clusters contained in one multiplet are topologically and genomic similar to each other. This is different from the research question tackled here, namely to detect the orthology signal at a functional complex level in a single PPI network given another species. The methodology proposed in this study generates orthology-driven clusters that contain evolutionary functional signal but are not in general conserved across species. Indeed, this could correspond to scenarios where a functional module retains its evolutionary origin while it changed its conformation in other species after speciation, due to evolutionary (pressure) events, resulting in a cluster with some links preserved, other being created and some links being lost.

The chapter is organized as follows. At first, Section 3.2 presents the data and methods used for our analysis. Next, the findings are discussed in Section 3.3. Finally, Section 3.4 provides the conclusions of the study.
the final yeast PPI network consists of 3545 proteins and 14354 interactions.

For obtaining orthology information we used the Inparanoid Database of Pairwise Ortholog\(^2\) (Berglund et al., 2008). This database contains clusters of ortholog groups (COGs) constructed by the Inparanoid program (Remm et al., 2001), which is a fully automatic method for finding orthologs and in-paralogs between two species. Ortholog clusters in the Inparanoid are seeded with a two-way best pairwise match (the seed ortholog pair), after which an algorithm for adding in-paralogs is applied. Because in-paralogs are homologs that arise when duplication occurs after speciation, and the duplicated gene often still retains the function of the ortholog (Dolinski and Botstein, 2007), they should be likely found in one protein complex. Therefore we consider all proteins present in COGs for inducing the PPI sub-network of orthologs and, for simplicity, we consider all proteins in a COG as orthologs. Specifically, in this study we call orthologous protein or ortholog a protein which is a part of an orthologous cluster produced by the Inparanoid when comparing two species.

In our analysis, COGs were obtained for the following pairs of organisms:

- *Saccharomyces cerevisiae* versus *Escherichia coli*,
- *Saccharomyces cerevisiae* versus *Caenorhabditis elegans*,
- *Saccharomyces cerevisiae* versus *Drosophila melanogaster*,
- *Saccharomyces cerevisiae* versus *Homo sapiens*.

*Escherichia coli* (E.coli), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly) and *Homo sapiens* (human) are standard organisms used in protein network and genome comparative studies (see e.g. Bhardwaj and Lu, 2005, Sharan et al., 2005b) and represent the diverse life-forms from a prokaryote (E.coli) to the highly complex eukaryote (human).

Yeast proteins in the derived ortholog groups are called yeast orthologs. Notice, each species comparison produces a different set of yeast orthologs.

\(^2\)http://inparanoid6.sbc.su.se
3.2. Methods

to be investigated. As a result, we consider the following 4 sets of yeast orthologs (present in the yeast PPI data), namely *yeast-E.coli*, *yeast-worm*, *yeast-fly*, *yeast-human*, consisting of 451, 1664, 1724, and 1850 proteins, respectively.

### 3.2.2 Quantifying orthology signal

We are interested in quantifying the orthology signal by means of a set of functions of those putative protein complexes detected by applying a clustering algorithm to a PPI network. To this end, we directly exploit evolutionary information of proteins as described by the presence of orthologs in another, given species. We call these proteins ‘true orthologs’. The following terminology is used in the sequel. A PPI network is represented by means of a graph \( G(V, E) \), where \( V \) is the set of nodes (proteins) and \( E \) is the set of edges (binary interactions). Let \( X \) be a subset of nodes \( V \) (e.g. ortholog set). The set \( X \) induces a sub-graph \( G[X] = (X, E_X) \) of \( G \), with set \( X \) of nodes and set \( E_X \) of those edges of \( E \) that join two nodes in \( X \). For a set \( S \), we denote by \( |S| \) the number of its elements.

Given a PPI network \( G = (V, E) \) and a given species \( s \), we propose a methodology for detecting the orthology signal at a functional complex level, consisting of the following steps.

1. **Retrieve from a database the set \( O \) of ‘true orthologs’ of \( V \) with respect to \( s \), with \( |O| = n \).**

2. **Generate the following three classes of clusters, using a given clustering algorithm.**

   (a) **Class 1 clusters (GC).** Apply clustering to the whole PPI network \( G \).

   (b) **Class 2 clusters (OC).** Apply clustering to the sub-network induced by \( O \).

   (c) **Class 3 clusters (RC).** Apply clustering to the sub-network induced by a randomly selected subset of \( V \) of size \( n \). Repeat the
process a number $N$ of times. Consider all sets of clusters detected across these runs ($RC = \{RC_1, RC_2, \ldots, RC_N\}$).

3. For each class of clusters,

(a) Infer putative complexes and identify their functions.

(b) For each identified function, compute its retrieval index as the fraction of those proteins in the detected complexes which have been assigned to that function and experimentally verified to have that function.

4. Select the set of those functions derived using putative complexes from class $OC$ and whose fractions are higher than those of the same function derived using putative complexes from class $GC$ and from class $RC$.

5. Output the set of putative complexes from class $OC$ having at least one of the selected functions.

The set of putative complexes of class $GC$ represent results of no selection (global or suppressor) bias and the collection of the sets of putative complexes present in the class $RC$ corresponds to the random selection bias. Accordingly, the complexes of class $OC$ represent the orthology selection bias. Thus, the method considers the complexes exhibiting orthology signal as those of the $OC$ class having a function which may not be attributed neither to the global bias nor to the random bias.

Next, we discuss the details of main steps of the proposed methodology.

**Generating the cluster classes**

Each class of clusters is produced by applying a clustering technique to the corresponding PPI (sub-)network. In this study we used the MCL clustering. MCL (van Dongen, 2000) computes clusters based on simulation of stochastic flow in graphs and it is widely used on many domains. It is able to use information on weights of edges of a given network if available. A first successful application of this algorithm on biological networks was
3.2. Methods

presented by Enright et al. (2002); MCL was also modified for detecting orthologous groups (Li et al., 2003). A recently published comparative study indicated that MCL outperforms other algorithms for clustering PPI networks (Brohee and van Helden, 2006). The inflation parameter of the algorithm was set to 1.8 as suggested by Brohee and van Helden (2006).

Inferring putative complexes and their functionalities

We want consider putative protein complexes containing more than a single protein-protein interaction. Therefore, after applying the clustering method we retain only clusters of size greater than or equal to 3. In order to infer the putative functions of a cluster, we measured the enrichment of functional annotations of the corresponding protein set, as entailed by the GO annotation (Ashburner et al., 2000), using one of the well-established tools, the Ontologizer\(^3\) (Bauer et al., 2008). The Ontologizer offers various algorithms for measuring GO enrichments. Here, we apply the standard statistical analysis method based on the one-sided Fisher’s exact test (Bauer et al., 2008), which measures the statistical significance of an enrichment and assigns to the cluster a p-value for each enriched function. The p-value is further corrected for multiple testing by means of a Bonferroni correction procedure.

The GO is known to have a hierarchical structure (directed acyclic graph) which can be used to define the level of an annotation. Specifically, the level of an annotation is equal to the length of the furthest path from the root of GO hierarchy to that annotation. This strategy always defines a filial annotation to have a higher level (deeper in the hierarchy) than its all parental annotations and hence no inconsistency on the description of GO hierarchical level (a parent having the same or higher level than its child) is introduced. The GO terms closer to the root of GO give more general description of biological functions while terms closer to the leaves of GO have granular and very specific biological definitions.

To measure functional annotation enrichments of proteins present in a cluster we used only experimentally verified annotations as reported in

\(^3\)http://compbio.charite.de/index.php/ontologizer2.html
the yeast gene association file of Saccharomyces Genome Database\(^4\) (SGD), available at the GO database\(^5\). We excluded all computationally assigned annotations to yeast proteins to avoid introducing a possible bias, because many of these techniques use protein structure or sequence similarity which may often refer to orthology.

Each detected cluster is a potential protein complex. The quality of a protein cluster is given by the coherence of biological functions of proteins contained in the cluster. If a certain subset of proteins in a cluster has a significantly coherent function, a prediction of that function for all proteins in the cluster can be made. Note that one may obtain more than one protein function prediction if more significantly coherent functions in the cluster are found. We say that proteins of a cluster have a *significantly coherent function* or functional GO annotation if the following criteria are satisfied:

1. the GO annotation is significantly enriched by the proteins in the cluster (p-value < 0.001).
2. more than half of the proteins in the cluster has this significant annotation.
3. the annotation is at least at the GO level four from the root of GO hierarchy.

In such a case the cluster can be used as protein function predictor and the significantly enriched GO annotation of the cluster is used to predict protein function of each of the proteins in that cluster. If a cluster does not satisfy the above conditions, no prediction can be made. Similar criteria were used by e.g Liang et al. (2006) or Jancura and Marchiori (2010). The condition on GO hierarchy guarantees that the prediction about biological functions is sufficiently specific and informative (Yon Rhee et al., 2008).

\(^4\)http://www.yeastgenome.org/
\(^5\)http://www.geneontology.org/GO.downloads.shtml

3.2. Methods

Each cluster which is a predictor defines a *putative protein complex* and the set of significantly coherent functions defines the set of inferred functions.

In the last step, for each putative complex we do an additional inference analogous to the protein function annotation procedure as follows. The GO hierarchy defines a parent-child relationship between GO functional terms where each descendant inherits all features of its ancestors. As a consequence, once a protein has a GO term annotation assigned, the protein has implicitly also annotations of all parental terms of the annotated function. Hence, using the same ratio, given the set of inferred functions of a putative protein complex, the complex also inherits all parental GO terms of the inferred functions. Thus, we may distinguish the following two sets of annotations, the most granular, filial, annotations, where no parent-child relationship between corresponding GO terms may be observed, and the set of all annotations which is the union of the filial annotations and all its ancestral annotations in GO hierarchy. Notice, by the definition, for a given complex all filial annotations are significantly coherent functions of the complex while the parental annotations need not be significantly enriched.

**Estimating the retrieval index of GO functions**

Having a set or class of putative protein complexes, one can quantify, at a fine-grained, protein level, a so called retrieval index of functions inferred by the protein complexes and defined as follows.

Consider a PPI network $G(V, E)$ and let $X \subseteq V$. Let $G[X] = (X, E_X)$ be the corresponding induced sub-graph of $G$ and $X_0 \subset X$ be the set of singletons in $G[X]$; that means there is no edge (interaction) in $E_X$ containing any of the proteins in $X_0$. We define the set of background proteins as $B(X) = \{X \setminus X_0\}$ and we denote $S(X)$ the set of all proteins contained in putative complexes discovered in $G[X]$. Additionally, let $C(f) \subseteq V$ be the set of candidate proteins for function $f$, that is, the set of all proteins having either experimentally annotated function $f$ or an experimentally annotated function that is a descendant function of $f$ in the GO hierarchy. Then let $P(f, U) = \{U \cap C(f)\}$ be the set of those proteins which have an
experimental evidence for the function \( f \) and are present in the set \( U \subseteq V \). We can define the retrieval index of a function \( f \) in \( X \) as the following fraction:

\[
\varrho(f, X) = \frac{|P(f, S(X))|}{|P(f, B(X))|}.
\] (3.1)

This fraction measures the retrieval of a given function \( f \) from a protein sample \( X \) by the set of putative complexes identified in \( X \). It can be viewed as an index measuring how likely a given function is present in a given set of putative complexes with respect to a given set of proteins.

Note that \( \varrho(f, V) \) corresponds to the retrieval index of \( f \) for the GC class, \( \varrho(f, O) \) corresponds to the retrieval index of \( f \) for the OC class, and \( \varrho(f, R_i) \) corresponds to the retrieval index of \( f \) for the \( RC_i \in RC \) class, where \( R_i \) is the random protein sample used at the run \( i \) when building the RC class.

**Identifying orthology-related functions and complexes**

We consider a function \( f \) to be related to the orthology signal if it satisfied two conditions: (a) it has a higher retrieval (at the level of putative protein complexes) in the set of orthologs than in the set of proteins of whole network and (b) it is unlikely to be retrieved when using random sampling. The second condition is formalized by comparing \( \varrho(f, O) \) with the 95th percentile of the set of retrieval indexes of \( f \) in the RC class. Specifically, for each function \( f \) from the GO hierarchy such that \( C(f) \cap B(O) \neq \emptyset \), we compute its functional retrieval indexes for GC, OC and the RC classes. Then, the function \( f \) is orthology-related iff

\[
\varrho(f, O) > \max\{\varrho(f, V), \varrho(f, R_{95\%})\},
\] (3.2)

where \( R_{95\%} \) is a random protein sample \( R_i \) such that \( \varrho(f, R_i) \) is the up 95th percentile of the all \( \varrho(f, R_1), \ldots, \varrho(f, R_N) \).

Finally, if a putative complex of the OC class has at least one orthology-related function, we consider that complex to be orthology-related.
3.3 Results and discussion

At first we report the data on generating cluster classes, that is how many clusters and putative complexes were extracted from interaction data for every possible class as given by orthology protein sets. Then we discuss how well putative complexes of \( OC \) class are functionally differentiated from the complexes of global or random protein selection. Finally, examples of orthology-related functions and complexes are presented.

3.3.1 Generating the cluster classes

As mentioned above, a state-of-the-art method for detecting communities in biological networks known as \( MCL \) (van Dongen, 2000) was used for clustering networks. \( MCL \) was applied to generate the following classes of clusters according to the given yeast ortholog sets:

- **OYC-E** - yeast clusters found using the sub-network induced by the yeast-E.coli ortholog set.

- **OYC-W** - yeast clusters found using the sub-network induced by the yeast-worm ortholog set.

- **OYC-F** - yeast clusters found using the sub-network induced by the yeast-fly ortholog set.

- **OYC-H** - yeast clusters found using the sub-network induced by the yeast-human ortholog set.

These groups are of the \( OC \) class and we generally refer to them by the common name OYC (ortholog yeast clusters).

The following classes of clusters were generated using random sampling:

- **RYC-E** - yeast clusters found using the sub-network induced by random sampled proteins of the same number as the number of proteins in the yeast-E.coli ortholog set.
Chapter 3. Quantifying the orthology signal in a PPI network

- RYC-W - yeast clusters found using the sub-network induced by random sampled proteins of the same number as the number of proteins in the yeast-worm ortholog set.

- RYC-F - yeast clusters found using the sub-network induced by random sampled proteins of the same number as the number of proteins in the yeast-fly ortholog set.

- RYC-H - yeast clusters found using the sub-network induced by random sampled proteins of the same number as the number of proteins in the yeast-human ortholog set.

These groups belong to the RC class and we generally refer to them by the common name RYC (random yeast clusters). For each of the four cases given above we performed 1000 runs. Recall that every run produces one particular RYC group. In order to compare these clusters with the GYC or OYC one, we consider the average values of RYC groups computed over all 1000 simulations according to a given ortholog set.

Finally, when MCL was applied to the whole yeast network, we get clusters of the above-mentioned GC class, and we refer to them by the name GYC (global yeast clusters).

Table 3.1 reports the number of GYC, OYC and RYC clusters identified by MCL, the number of functional complexes extracted from these clusters, the average size of the clusters and of the complexes, and the percentage of the clusters that correspond to functional complexes. The number of clusters, complexes and their average size are similar for RYC and OYC, while, as expected, more clusters (hence complexes) are generated by using GYC, and their average size is slightly bigger than that of those detected using RYC and OYC. It is interesting to note that the sensitivity of the method does not change significantly with respect to the (sub-)network it is applied to, resulting in about 40% of the detected clusters being functional complexes for GYC, and in the range 37 – 44% for RYC and OYC. This indicates the robustness of MCL with respect to the considered sampling strategies.
3.3. Results and discussion

| Clust. Gr. | #Clusters | $|C|$ | #Complexes | $|C'|$ | Ratio |
|-----------|-----------|------|------------|------|-------|
| GYC       | 365       | 8    | 147        | 11.1 | 40.3% |
| OYC-E     | 37        | 5    | 14         | 5.5  | 37.8% |
| RYC-E     | 34.31 ($\pm$3.82) | 4.17 ($\pm$0.40) | 12.69 ($\pm$2.96) | 5.06 ($\pm$0.68) | 37.0% |
| OYC-W     | 181       | 7    | 80         | 8.7  | 44.2% |
| RYC-W     | 175.22 ($\pm$7.21) | 6.08 ($\pm$0.28) | 67.85 ($\pm$5.87) | 8.07 ($\pm$0.52) | 38.7% |
| OYC-F     | 191       | 7    | 80         | 9.33 | 41.9% |
| RYC-F     | 181.97 ($\pm$7.51) | 6.15 ($\pm$0.36) | 70.32 ($\pm$6.01) | 8.19 ($\pm$0.53) | 38.6% |
| OYC-H     | 203       | 7    | 90         | 9.33 | 44.3% |
| RYC-H     | 196.38 ($\pm$7.80) | 6.29 ($\pm$0.45) | 75.71 ($\pm$6.21) | 8.41 ($\pm$0.54) | 38.6% |

Clust. Gr. - the cluster group (class), #Clusters - the number of clusters, #Complexes - the number of complexes, $|C|$ - the average cluster or complex size, Ratio - the percentage of clusters that are functional complexes.

Table 3.1: The number and average size of detected protein clusters and putative protein complexes.

3.3.2 Identifying orthology signal at protein complex level

The detected putative complexes are used to identify orthology-related functions. For each class of putative complexes we compute the functional retrieval indexes with respect to the protein sample set from which the complexes were derived by applying the above-defined formula (3.1) (see Section 3.2.2). Then, for each function $f$ associated with complexes of the OYC class, we compare its retrieval index $\varrho(f, O)$ with the retrieval indexes $\varrho(f, R_{95\%})$ and $\varrho(f, V)$ for both RYC and GYC classes using the rule (3.2) (see Section 3.2.2).

One may consider the comparison of $\varrho(f, O)$ with $\varrho(f, R_{95\%})$ as the random sample filter and $\varrho(f, O)$ with $\varrho(f, V)$ as the global sample filter. Only if $\varrho(f, O)$ is greater than both $\varrho(f, R_{95\%})$ and $\varrho(f, V)$, then the function $f$ and with it associated OYC complexes are considered to be orthology-related. Application of the random sample filter differentiates the functions of the OYC class complexes from those which are likely to be observed within the complexes of class RYC and thus by chance. In the case of global sample filter it extracts functions which have greater retrieval rates within the complexes of class OYC than within the complexes of class GYC.
Hence, these functions are suppressed when considering complexes present in the global topology of the PPI network and are unveiled when considering only the complexes formed by orthologs.

Table 3.2 reports the effect of these filters on the number of functions and associated complexes of the OYC class, when they are applied separately and when they are combined. We may observe that the global sample filter has no reduction effect on the number of complexes although from about one third to one quarter of all functions are omitted. This substantiates that indeed the complexes consisting of orthologs are well-differentiated from the complexes observed in the global topology of the PPI network.

<table>
<thead>
<tr>
<th>Clust. Gr.</th>
<th>#Complexes</th>
<th>Total</th>
<th>$\varrho_R$</th>
<th>$\varrho_G$</th>
<th>$\max{\varrho_R, \varrho_G}$</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>OYC-E</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>251</td>
<td>150</td>
<td>184</td>
<td>144</td>
<td></td>
<td>57.37%</td>
</tr>
<tr>
<td>OYC-W</td>
<td>80</td>
<td>65</td>
<td>80</td>
<td>65</td>
<td>123</td>
<td>81.25%</td>
</tr>
<tr>
<td></td>
<td>767</td>
<td>124</td>
<td>526</td>
<td>123</td>
<td></td>
<td>16.04%</td>
</tr>
<tr>
<td>OYC-F</td>
<td>80</td>
<td>68</td>
<td>80</td>
<td>68</td>
<td>109</td>
<td>85.00%</td>
</tr>
<tr>
<td></td>
<td>775</td>
<td>109</td>
<td>487</td>
<td>109</td>
<td></td>
<td>14.06%</td>
</tr>
<tr>
<td>OYC-H</td>
<td>90</td>
<td>89</td>
<td>90</td>
<td>89</td>
<td>78</td>
<td>98.89%</td>
</tr>
<tr>
<td></td>
<td>735</td>
<td>79</td>
<td>444</td>
<td>78</td>
<td></td>
<td>10.61%</td>
</tr>
</tbody>
</table>

Clust. Gr. - the cluster group (class), #Complexes - the number of complexes, #Functions - the number of functions, Total - the numbers in total, $\varrho_R = \varrho(f, R_{95\%})$ - the numbers after applying the random sample filter, $\varrho_G = \varrho(f, V)$ - the numbers after applying the global sample filter, $\max\{\varrho_R, \varrho_G\}$ - the numbers after applying the both filters, Ratio - the percentage of complexes or functions which passes through the both filters.

Table 3.2: The effect of filtering procedures.

In the case of the random sample filter the number of functions drops considerably, whereas the number of complexes still remains high. As a result, when both filters are combined, one may interpret the total reduction on the number of complexes and functions as primarily caused by the random sample filter, while there is almost no effect on the reduction due to the global sample filter, especially on the number of complexes.

This suggests that in the set of all annotations associated with a given complex of the OYC class, it is very likely to observe an orthology-related
3.3. Results and discussion

<table>
<thead>
<tr>
<th>Clust. Gr.</th>
<th>GO ID</th>
<th>Name</th>
<th>GO Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>OYC-E</td>
<td>GO:0005840</td>
<td>ribosome</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td>GO:0005694</td>
<td>chromosome</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td>GO:0000228</td>
<td>nuclear chromosome</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0006811</td>
<td>ion transport</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0006519</td>
<td>cellular amino acid metabolic process</td>
<td>BP</td>
</tr>
</tbody>
</table>

Clust. Gr. - the cluster group (class), CC - cellular component, MF - molecular function, BP - biological process.

Table 3.3: Orthology-related functional categories for yeast-E.coli orthologs.

function despite the sparse distribution of orthology-related functions in the GO hierarchy. As a result, more than 80% of the OYC complexes are always indeed orthology-related complexes, which suggests they mostly do not correspond to an outcome of a stochastic event.

We discuss in the sequel some interesting orthology-related functions as well as novel protein function predictions derived using the proposed methodology.

3.3.3 On orthology-related functions

In the set of yeast orthologs with respect to E.coli we identified 144 orthology-related functions. Table 3.3 reports only higher level functions in GO hierarchy as determined by the GO slim functional terms. Each GO slim characterizes a certain type of biological functions which have some features and tasks in common, and hence they define the functional categories in a biological system.

Considering cellular compartments of a cell, we identified ribosomal and chromosomal proteins as being orthology-related. Indeed, it has been shown

6http://www.geneontology.org/GO.downloads.shtml
GO slim version: 1.1.1543, date: 19/10/2010.
Chapter 3. Quantifying the orthology signal in a PPI network

<table>
<thead>
<tr>
<th>Clust. Gr.</th>
<th>GO ID</th>
<th>Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GO:0042555</td>
<td>MCM complex</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td>GO:0004672</td>
<td>protein kinase activity</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:0004674</td>
<td>protein serine/threonine kinase activity</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:003883</td>
<td>CTP synthase activity</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:0043565</td>
<td>sequence-specific DNA binding</td>
<td>MF</td>
</tr>
<tr>
<td></td>
<td>GO:0009987</td>
<td>cellular process</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0044257</td>
<td>cellular protein catabolic process</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0051603</td>
<td>proteolys. in cell. prot. catab. proc.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0019941</td>
<td>modif.-dependent prot. catab. proc.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0006511</td>
<td>ubiquitin-dependent prot. catab. proc.</td>
<td>BP</td>
</tr>
<tr>
<td>OYC-W</td>
<td>GO:0006220</td>
<td>pyrimid. nucleotide metab. process</td>
<td>BP</td>
</tr>
<tr>
<td>OYC-F</td>
<td>GO:0009147</td>
<td>pyrimid. nucleos. triphosph. metab. pr.</td>
<td>BP</td>
</tr>
<tr>
<td>OYC-H</td>
<td>GO:0006221</td>
<td>pyrimid. nucleotide biosynthetic proc.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0009218</td>
<td>pyrimid. ribonucleotide metabolic proc.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0009208</td>
<td>pyrimid. ribonucleos. triphos. metab. pr.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0009148</td>
<td>pyrimid. nucleoside triphos. biosynth. pr.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0009220</td>
<td>pyrimid. ribonucleot. biosynth. proc.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0009209</td>
<td>pyrimid. ribonucleos. triphos. biosyn. pr.</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0046036</td>
<td>CTP metabolic process</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>GO:0006241</td>
<td>CTP biosynthetic process</td>
<td>BP</td>
</tr>
</tbody>
</table>

Clust. Gr. - the cluster group (class), CC - cellular component, MF - molecular function, BP - biological process.

Table 3.4: Orthology-related functions for yeast-worm, yeast-fly, and yeast-human orthologs.

that the ribosomes in the mitochondria of eukaryotic cells resemble those in bacteria, reflecting the likely evolutionary origin of this organelle (Benne and Sloof, 1987). Considering other reported functional categories, numerous phylogenetic data provide strong evidences that there is a constant evolutionary pressure in conserving critical functional domains on proteins that are significant for cell survival. These proteins are usually components of DNA/RNA replication, transcription and translation apparatus or they are involved in ion transport processes.

Because worm, fly and human all belong to eukaryotes, we looked at their common orthology-related functions (reported in Table 3.4). Con-
sidering molecular functions, we retained mostly kinases activity proteins and DNA binding proteins. This is true in particular for proteins of kinase activity, which have been found conserved among eukaryotes: these kinase’ functional conservations were investigated for yeast, worm, fly and human when studying their evolution (Manning et al., 2002). Orthology-related DNA binding proteins have been also known to exhibit high sequence conservation among eukaryotes (e.g. Brandt et al., 2009, Chang et al., 2008).

Regarding the Mcm complex, it consists of six eukaryotic Mcm proteins which also share significant sequence similarity with one another. These proteins serve as the eukaryotic replicative helicase, the molecular motor that both unwinds duplex DNA and powers fork progression during DNA replication (Bochman and Schwacha, 2009) and therefore are expected to be orthology-related.

CTP and pyrimidine processes are incorporated in the growth of RNA and DNA during the process of transcription or DNA replication. Short-term energy storage is also one of the functions of pyrimidines. Hence, as mentioned above, there is a pressure on evolutionary conservation of these processes vital for a cell survival. Last but not least, proteins involved in ubiquitination-dependent processes contain a highly conserved ubiquitin-conjugating (UBC) domain; thus, the function is also orthology-related.

### 3.3.4 On orthology-related complexes and novel predictions

Orthology-related complexes are those complexes of the OYC class whose proteins perform at least one orthology-related function. In addition, we call unique complexes those complexes whose proteins have a predicted function that is not inferred for those proteins by any GYC complex. These are the complexes that are new and derived using (the protein complex composition present in) the orthology sub-network, that is, uniquely linked to the orthology signal.

Given a unique cluster and its protein having a new predicted function not inferred by any GYC complex containing the protein. Then, if the function prediction is experimentally or computationally annotated in SGD, this prediction is verified. Analogously, if we find the new predicted function
has not been experimentally or computationally annotated in SGD, then this prediction is indeed a novel prediction. Observe that one cluster can have verified as well as novel predictions at the same time. The number of orthology-related complexes as well as those which produce verified and/or novel protein function predictions are reported in Table 3.5. We may observe that, for each ortholog set, from all complexes with a novel prediction, more than 80% are orthology-related complexes. This is analogous to the reduction effect on the whole set of complexes mentioned above.

Examples of novel orthology-related complexes are given in Table 3.6: they demonstrate that by examining different sets of orthologs we found specific putative complexes, most of them crucial for a living cell.

For instance, proteins of Cluster 1 are predicted to be involved in mitochondrial proton-transporting ATP synthase, catalytic core. While ATP1 and ATP2 are indeed the part of the catalytic core, ATP3 is part of the central stalk of mitochondrial proton-transporting ATP synthase. Cluster 1, however, gives a proper suggestion for the mechanism of the ATP3. Moreover, as ATP3 interacts with ATP2 it may be involved also in the catalytic core.

<table>
<thead>
<tr>
<th>Clust. Gr.</th>
<th>Total</th>
<th>Unique</th>
<th>Verified</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>OYC-E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#All</td>
<td>14</td>
<td>13</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>#Ort.-related</td>
<td>14</td>
<td>13</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>OYC-W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#All</td>
<td>80</td>
<td>69</td>
<td>12</td>
<td>63</td>
</tr>
<tr>
<td>#Ort.-related</td>
<td>65</td>
<td>57</td>
<td>11</td>
<td>52</td>
</tr>
<tr>
<td>OYC-F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#All</td>
<td>80</td>
<td>62</td>
<td>12</td>
<td>59</td>
</tr>
<tr>
<td>#Ort.-related</td>
<td>68</td>
<td>52</td>
<td>11</td>
<td>50</td>
</tr>
<tr>
<td>OYC-H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#All</td>
<td>90</td>
<td>72</td>
<td>10</td>
<td>68</td>
</tr>
<tr>
<td>#Ort.-related</td>
<td>89</td>
<td>72</td>
<td>10</td>
<td>68</td>
</tr>
</tbody>
</table>

Clust. Gr. - the cluster group (class), #All - the number of all complexes, #Ort.-related - the number of orthology-related complexes.

Table 3.5: The numbers of putative protein complexes containing unique, verified and novel protein function predictions.
### 3.3. Results and discussion

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Proteins</th>
<th>Prediction</th>
<th>Cluster Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cluster 1</strong></td>
<td>ATP1</td>
<td>mitochondrial</td>
<td>OYC-E</td>
</tr>
<tr>
<td></td>
<td>ATP2</td>
<td>proton-transporting ATP synthase, catalytic core</td>
<td>OYC-E</td>
</tr>
<tr>
<td></td>
<td>ATP3</td>
<td>OYC-E</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 2</strong></td>
<td>MMS2</td>
<td>ubiquitin conjugating enzyme complex</td>
<td>OYC-W, OYC-F, OYC-H</td>
</tr>
<tr>
<td></td>
<td>UBC13</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ERR3</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 3</strong></td>
<td>UBC7</td>
<td>protein ubiquitination</td>
<td>OYC-W, OYC-F, OYC-H</td>
</tr>
<tr>
<td></td>
<td>UBC5</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UBC6</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UBC1</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UBC8</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UBC4</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VIP1</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 4</strong></td>
<td>SEC22</td>
<td>Golgi vesicle transport, Golgi apparatus</td>
<td>OYC-W, OYC-F, OYC-H</td>
</tr>
<tr>
<td></td>
<td>SFT2</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SED5</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BET3</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SLY1</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEC17</td>
<td>OYC-W, OYC-F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEC18</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COS1</td>
<td>OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SYM2</td>
<td>OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GPA1</td>
<td>OYC-W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STE4</td>
<td>OYC-W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AKR1</td>
<td>OYC-W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YKT6</td>
<td>OYC-W</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 5</strong></td>
<td>SEC23</td>
<td>COPII vesicle coat</td>
<td>OYC-W, OYC-F, OYC-H</td>
</tr>
<tr>
<td></td>
<td>SEC24</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SFB2</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIP1</td>
<td>OYC-W, OYC-F, OYC-H</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: (continued at the next page)
Chapter 3. Quantifying the orthology signal in a PPI network

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Proteins</th>
<th>Prediction</th>
<th>Cluster Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GRH1</td>
<td></td>
<td>OYC-W,OYC-F</td>
</tr>
<tr>
<td></td>
<td>BUG1</td>
<td></td>
<td>OYC-F</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>SEC9</td>
<td></td>
<td>OYC-W,OYC-H</td>
</tr>
<tr>
<td></td>
<td>SNC1</td>
<td></td>
<td>OYC-W,OYC-H</td>
</tr>
<tr>
<td></td>
<td>SNC2</td>
<td></td>
<td>OYC-W,OYC-H</td>
</tr>
<tr>
<td></td>
<td>SSO1</td>
<td></td>
<td>OYC-W,OYC-H</td>
</tr>
<tr>
<td></td>
<td>SSO2</td>
<td></td>
<td>OYC-W,OYC-H</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>ATG5</td>
<td></td>
<td>OYC-F</td>
</tr>
<tr>
<td></td>
<td>ATG7</td>
<td></td>
<td>OYC-F</td>
</tr>
<tr>
<td></td>
<td>SSO2</td>
<td></td>
<td>OYC-F</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>HXT3</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>HXT2</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>HXT4</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>HXT1</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>SNF3</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>RGT2</td>
<td></td>
<td>OYC-E,OYC-F</td>
</tr>
<tr>
<td></td>
<td>CYC8</td>
<td></td>
<td>OYC-E</td>
</tr>
</tbody>
</table>

Table 3.6: Novel orthology-related complexes.

Cluster 2 and 3 are ubiquitin complexes. In general, in eukaryotes ubiquitin-dependent processes relate to protein degradation, because it is catalysed by a family of ubiquitin-carrier enzymes (E2) which contain a highly conserved ubiquitin-conjugating (UBC) domain. Previous reports showed that numerous members of this family are functionally overlapping (Seufert and Jentsch, 1990, van Wijk and Timmers, 2010). Hence, as one could expect, our complexes, Cluster 2 and Cluster 3, are found for all eukaryotic yeast’s orthologs, consisting of ubiquitin-conjugating enzymes that mediate protein degradation, indicating a highly conservation of UBCs during evolution for eukaryotes. In Cluster 2 ERR3 is a protein of unknown function, which has similarity to enolases. This suggests that ERR3 is part of the ubiquitin conjugating enzyme complex. In case of Cluster 3
the VIP1 was the only protein found with no UBC activity indicating the involvement of kinases in the complex process of ubiquitination. However, experimental data demonstrated the ubiquitin-proteasome machinery to control the levels of kinases by proteolysis (Lu and Hunter, 2009). As the mechanism of ubiquitin-mediated protein degradation is poorly understood it requires further investigation.

Next, we discuss proteins in the closely related complexes Cluster 4 and Cluster 5. The protein families that mediate vesicle trafficking are conserved through phylogeny from yeast to human, as well as throughout the cell from the endoplasmic reticulum to the plasma membrane (Bock et al., 2001). Our analysis showed proteins of the SEC family (SEC22, SEC23 and SEC24) and others as SED5, BET3, SLY1, HIP1 and SFB2 conserved from worm to human and involved in the coat protein complex II (COPII) that selectively transport molecules and vesicle fusion proteins from the endoplasmic reticulum (ER) to the Golgi complex (Kuehn et al., 1998). Other proteins included in the complexes but not for all species as BUG1 and GRH1 have been observed co-localizing on the cis-Golgi and they form a heterooligomeric complex binding GRH1 at the well conserved C terminus of BUG1 (Behnia et al., 2007). The role of these two proteins in ER to Golgi transport is mediated by the interaction between GRH1 with the SEC23/24 complex, proteins that we could identify in the same complex a that of GRH1 and BUG1. In these clusters related with vesicle trafficking we could observe other proteins like SFT2, COY1 and GOS1 not annotated for the ER to Golgi vesicle-mediated transport term but we could classify them in the correct cluster. These proteins have been observed to be required for vesicle fusion with the Golgi complex (Conchon et al., 1999, Gillingham et al., 2002).

Further interesting outcomes are Cluster 6 and Cluster 7. Both clusters share the SSO2 protein but they produce different functional predictions. In the case of Cluster 6 SSO2 interacts with proteins of the yeast SNARE complex (SEC9, SNC1, SNC2, SSO1), the core of the machinery required for membrane fusion, while in Cluster 7 SSO2 is involved with the Cvt pathway proteins (ATG5, ATG7), a biosynthetic transport route for a distinct subset of resident yeast vacuolar hydrolases. Reggiori et al. (2004)
Chapter 3. Quantifying the orthology signal in a PPI network

described that the biogenesis of Cvt vesicles apparently requires a fusion step catalysed by the VFT tethering factor and by the SNARE complex but they failed to show the proteins that are related in the interaction between the Cvt pathway and the SNARE complex. Although to further elucidate the real role of the SSO2 protein in the interaction between the Cvt pathway and the SNARE complex an experimental validation is necessary, these results show the capability of the presented methodology not only to classify proteins interacting within the same or related clusters but also to predict unknown protein interactions between different pathways and complexes that are currently under investigation. Interestingly, Cluster 6 is found in OYC-W and OYC-H cluster groups corresponding to yeast-worm and yeast-human ortholog sets while Cluster 7 is found only in OYC-F corresponding to the yeast-fly ortholog set suggesting the complexity and versatility of protein complex evolution.

Finally, we discuss Cluster 8, which contains the novel prediction for the SNF3 and RGT2 proteins. This complex was observe for yeast-fly and yeast-E.coli ortholog sets but not for the other ones. Previous studies in yeast demonstrated that SNF3 and RGT2 are integral membrane proteins with unusually long carboxy-terminal tails involved in glucose transport. This is in compliance with our results that showed both proteins to have a glucose transport activity. However, according to recent studies, although both proteins are very similar to glucose transporters, they apparently do not transport glucose but they interact as glucose sensors. Özcan et al. (1998) demonstrated that glucose signalling is not the result of glucose transport and that the C-termini of both proteins are signalling domains of these glucose sensors. Nonetheless, it remains unclear how glucose transport is regulated and therefore our prediction can be considered as valid. In addition, as the SNF3/FGT2 protein interaction was not found in yeast ortholog set with respect to human and worm, it indicates that the protein complex is not conserved among all species. Aside the SNF3/RGT2 complex the predicted cluster includes also the HXT-transporters which are responsible for glucose uptake. Moreover, in OYC-E this protein complex was assembled with the contribution of CYC8, a yeast protein that binds to the promotors of the HXT genes blocking their transcription. This finding
is very interesting as E.coli contains no nucleus and therefore it is likely that an equivalent protein complex exists.

3.4 Conclusions

We proposed a novel methodology for quantifying the functionality of the orthology signal in a PPI network at a functional complex level. The methodology performs a differential analysis between the functions of those complexes detected by clustering a PPI network using only proteins with orthologs in another given species, and the functions of complexes detected using the entire network or sub-networks generated by random sampling of proteins.

Results of our experimental analysis indicated the usefulness of the proposed methodology to identify functional categories and complexes that can be clearly attributed to the presence of an evolutionary (orthology) signal, as supported by biological evidence from related studies.

As a future work, we intend to investigate possible extension of the methodology to increase its sensitivity. In particular one can exploit the inheritance property present within the GO hierarchy, namely each filial GO term may inherit features of its parental terms. For example, one could propagate the evolutionary signal between the two closest orthology-related function in the GO hierarchy such that all GO terms present on the paths between these two terms are also orthology-related.
Chapter 4

Dividing protein interaction networks for detection of conserved complexes

A protein complex observed across different species is assumed to originate from a common ancestor throughout evolution. Analysing PPI networks in order to identify such evolutionary conserved complexes is important for understanding how PPI networks were shaped by evolutionary processes. Here, we present an algorithm for dividing PPI networks into small sub-graphs that are likely to cover conserved complexes. The proposed dividing algorithm combines a graph theoretical property (articulation) with a biological property (orthology). Extensive experiments on various PPI networks are conducted in order to assess how well the sub-graphs generated by this dividing algorithm cover evolutionary conserved protein complexes.1

1This chapter is based on the following works:
Chapter 4. Dividing protein interaction networks

4.1 Introduction

With the exponential increase of data on protein interactions obtained from advanced technologies, data on thousands of interactions in human and most model species have become available (e.g. Bader et al., 2001, Xenarios et al., 2002). PPI networks offer a powerful representation for better understanding modular organization of cells, for predicting biological functions and for providing insight into a variety of biochemical processes.

Recent studies consider a comparative approach for the analysis of PPI networks from different species in order to discover common protein groups, called conserved complexes, which are likely to be related and to share similar functionality in a cell (Sharan and Ideker, 2006, Srinivasan et al., 2007). This problem is known as protein network alignment and many methods for network alignment have been proposed (Jancura and Marchiori, 2012).

The aim of this chapter, however, is not to propose yet another network alignment algorithm, but to show how PPI networks can be divided into small sub-graphs that are likely to cover conserved complexes. Such method can be used in connection with a network alignment, as a prior pre-processing step, in order to enhance the computational strategy or to parallelize the alignment computation.

One can observe certain properties related with conserved complexes. For instance, conserved complexes discovered by computational techniques have in general small size (that is, number of proteins) compared to the size of the PPI network they belong to. Moreover, PPI networks are known to have a scale-free topology where most proteins participate in a small number of interactions while a few proteins, called hubs, contain a high number of interactions. As indicated by a recent study, hubs whose removal disconnects a PPI network (articulation hubs) are likely to appear in conserved interaction patterns (Pržulj, 2005).

These observations motivate the introduction of an algorithm for dividing PPI networks, called Divide, that combines biological (orthology) and graph theoretical (articulation) information: it detects small groups of ortholog articulations, called centers, which are then expanded into subsets of ortholog nodes. This algorithm has the desirable property of being
parameterless.

The effectiveness and robustness of Divide is assessed experimentally in the following ways.

First, we show that the sub-graphs generated by Divide indeed cover "true" conserved protein complexes. This is done by measuring the overlap of these sub-graphs with MIPS curated functional complexes restricted to those proteins belonging to an orthologous pair.

Secondly, we show that the generated sub-graphs cover protein complexes computationally predicted. Specifically, we compare these sub-graphs with the conserved complexes predicted by one state-of-the-art pairwise local alignment algorithm, called MaWish (Koyutürk et al., 2006b). We investigate experimentally how the application of Divide retains the complexes detected by MaWish, and whether the generated sub-graphs contain information to be used for discovering new conserved complexes. Results of an extensive experimental analysis indicate that indeed Divide generates sub-graphs containing conserved complexes that are not detected by MaWish.

To the best of our knowledge, we propose the first algorithm that tackles the issue of identifying sub-graphs that cover conserved complexes. In contrast to standard clustering methods applied to PPI networks or alignment methods, the Divide sub-graphs are not putative ad-hoc (conserved) protein complexes but rather contain one or multiple complexes to be still extracted. Moreover, the Divide algorithm is independent of any alignment technique.

In general, the results of this study substantiate the important role of orthologous articulations in evolutionary analysis of PPI networks and their possible application for detection of conserved protein complexes.

This chapter is organized as follows. In the next section we discuss related works. Section 4.3 describes the graph-theoretic terminology used in the chapter. The Divide algorithm is introduced in Section 4.4. Section 4.5 summarizes the data and the type of assessment employed in the experimental analysis. In Section 4.6 the robustness of Divide is assessed by analysing how the generated sub-graphs cover "true" complexes. In Section 4.7 the sub-graphs generated by Divide are compared with the complexes
Chapter 4. Dividing protein interaction networks

predicted by MaWish. Finally, we conclude and briefly address future works in Section 4.8.

4.2 Related work

The importance of highly connected proteins (hubs) within a PPI network has been widely investigated (e.g. Ekman et al., 2006, Jeong et al., 2001, Pržulj, 2005, Pržulj et al., 2004, Ucar et al., 2006). In particular, it was shown by Jeong et al. (2001) that hubs with a central role in the network architecture are three times more likely to be essential than proteins with only a small number of links to other proteins. In addition, further studies revealed hubs are more evolutionary conserved than non-hub proteins (Fraser et al., 2002) and that different types of hubs may exhibit different evolutionary rates (Aragues et al., 2007, Fraser, 2005). These initial premises have been later corroborated by major literature evidence as recently surveyed by Jancura and Marchiori (2012).

The knowledge on hub existence and their functional roles has been exploited by various researchers for data-mining PPI networks. For example, Lee et al. (2009) used hub prioritization to rank proteins for the identification of novel components of the midbody proteome. Next, Ucar et al. (2006) proposed a network refinement method based on a hub duplication in order to identify proper functional groupings of proteins. Hwang et al. (2009) built a machine learning classifier to predict essential genes from topological as well as sequence features, including the connectivity of a protein and its evolutionary conservation.

The tendency of hubs to evolve slowly was also used for the purpose of detecting evolutionary conserved complexes (Gerke et al., 2007). Specifically, Gerke et al. (2007) avoid all-against-all protein sequence comparison between two species by identifying hubs and clusters of proteins in the corresponding PPI networks prior the sequence alignment. Then, putative orthologous proteins are identified by sequence comparison only between the proteins contained by the clusters and regions centred around the hubs. Finally, pairs of network regions between species are established
4.3 Graph theoretic background

and their topological matchings are performed in order to determine evolutionary conserved complexes. The focus of this method was on reducing the computational time needed for extensive protein sequence alignment. Hence, the topology of the network was exploited without knowing putative orthologous proteins.

The aim of this study is not to reduce the number of sequence alignments to be performed in order to compare PPI networks, but, in contrast, to incorporate the knowledge on orthology as established by sequence alignments to efficiently divide PPI networks considering only putative orthologs. Thus, our method combines topological as well as biological properties of proteins in PPI networks. Moreover, the hubs and clusters constructed by the method in Gerke et al. (2007) is primarily devised for the consecutive network alignment model, which may still locally exploit the PPI network around the hubs, while here we propose a divisive algorithm independent of any network comparison model such that the resulting sub-networks contain conserved protein complexes.

In general, there is no consensus on the definition of hubs in a PPI network. Mostly, the topological measure on the number of node’s interactions, called the degree of a node, is employed, where a certain cut-off value is used to distinguish hubs (high degree nodes), and non-hub proteins (low degree nodes). This study avoids the problem of tuning such threshold parameter by relying on the following observation. Pržulj (2005) discovered that if one takes functional groups in PPI networks, then, amongst all functional groups, cellular organization proteins have the largest presence in those hubs whose removal disconnects the network. This type of hubs are called articulation hubs and the above-mentioned woks justify the use of orthologous articulation hubs for dividing PPI networks in order to identify regions containing functional conserved protein complexes.

4.3 Graph theoretic background

Given a graph $G = (U, E)$, nodes joined by an edge are called adjacent. A neighbor of a node $u$ is a node adjacent to $u$. The degree of $u$ is the number
of elements in $E$ containing the vertex $u$.

A graph $G = (U, E)$ is called undirected if $uu'$ in $E$ implies $u'u$ also in $E$; otherwise $G$ is called directed. A directed acyclic graph is a directed graph that contains no cycles.

A sub-graph $H(V, F)$ of an undirected graph $G(U, E)$ is said to be induced by the set of nodes $V \subseteq U$ if and only if the set of edges $F \subseteq E$ consists of all the edges that appear in $G$ over the same vertex set $V$.

A graph is connected if there is a path from any node to any other node. Let $G(U, E)$ be a connected undirected graph. A vertex $u \in U$ is called articulation if the graph resulting by removing this vertex from $G$ and all its edges, is not connected.

A tree is a connected graph not containing any circle. A tree is called rooted tree if one vertex of the tree has been designated as the root. Given a rooted tree $T(V, F)$, the depth of a vertex $v \in V$ is the number of edges from the root to $v$ without repetition of edges. Leaves of the tree $T$ are vertices which have only one neighbor. The depth of a tree is the highest depth of its leaves. A spanning tree $T(V, F)$ of a connected undirected graph $G(U, E)$ is a tree where $V = U$ and $F \subseteq E$.

Given an edge-weighted (or node-weighted) graph $G(U, E)$ with a scoring function $w : e \in E \rightarrow \mathbb{R}$ (or $w : u \in U \rightarrow \mathbb{R}$). Total weight $w(G)$ of $G$ is the sum of weights of all edges (or nodes) in the graph:

$$w(G) = \sum_{e \in E} w(e) \quad (\text{or } w(G) = \sum_{u \in U} w(u)).$$

Suppose a connected undirected graph $G(U, E)$ and a vertex $u \in U$ are given. Let $N(u)$ a set of all neighbors of $u$ and $N'(u) \subseteq N(u)$ be. A center of $u$ is the set $C(u) \equiv N'(u) \cup \{u\}$.

Observe that a center can be expanded to a spanning tree of $G(U, E)$. Moreover, the center as an initial set of expansion can be consider as a root if we merge all vertices of center to one node. Such spanning tree created from a center, called centered tree, has zero depth all vertices of center and the vertices of $i$-depth are new nodes added in $i$th iteration of expansion to the spanning tree. Therefore a centered tree, can be generated as follows:
4.3. Graph theoretic background

- the 0-depth of the centered tree is the center

- the $i$-th depth of the centered tree consists of all neighbors of $(i-1)$-th depth which are not yet in any lower depth of the centered tree yet.

Examples of a spanning and centered tree are shown in Figure 4.1.

![Spanning tree](image1)

![Centered tree](image2)

Figure 4.1: Examples of spanning and centered tree in the same graph. The dark grey node in the left figure represents a root. Dark grey nodes in the right figure represent a center. Numbers indicate depths of nodes in trees. Solid edges are edges of a spanning tree. Dash edges are other edges of the graph.

A PPI network is represented by an undirected graph $G(U, E)$. $U$ denotes the set of proteins and $E$ denotes set of edges, where an edge $uu' \in E$ represents the interaction between $u \in U$ and $u' \in U$. Given PPI networks $G(U, E)$ and $H(V, F)$. A vertex $u \in U$ is orthologous if there exists at least one vertex $v \in V$ such that $uv$ is an orthologous pair. Orthologous articulation is an orthologous vertex which is an articulation. An orthology path is a path containing only orthologous vertices.
4.4 Divide algorithm

Suppose given the PPI networks $G$ and $G_1$ of two species. Let $G(U, E)$ and $O \subseteq U$ be the set of vertices which are orthologous w.r.t. the vertices of $G_1$. Suppose $O$ contains $n$ elements. The Divide algorithm is shown in pseudo-code in Algorithm 1. It generates centers from orthologous articulations and expands them into centered sub-trees containing only orthologous proteins. The main steps of Divide are described in detail below.

**Computing Articulations** (Line 1). Computation of articulations can be performed in linear time by using, e.g., Tarjan’s algorithm described in Tarjan (1972) or Hopcroft and Tarjan (1973).

**Greedy Construction of Centers** (Lines 3-10). The degree (in $G$) of all orthologous articulations is used for selecting seeds for the construction of centers. Networks with scale-free topology appear to have edges between hubs systematically suppressed, while those between a hub and a low-connected protein seem favoured (Maslov and Sneppen, 2002). Guided by this observation, we greedily construct centers by joining one orthologous articulation hub with its orthologous articulation neighbors, which will more likely have low degree.

Specifically, let $A$ be the set of orthologous articulations of $G$. The first center consists of the element of $A$ with highest degree and all its neighbors in $A$. The other centers are generated iteratively by considering, at each iteration, the element of $A$ with highest degree among those which do not occur in any of the centers constructed so far, together with all its neighbors in $A$ which do not already occur in any other center. The process terminates when all elements of $A$ are in at least one center. Then an unambiguous label is assigned to each center.

**Initial Expansion** (Lines 11-16). By construction, centers cover all orthologous articulations. Articulation hubs are often present in conserved sub-graphs detected by means of comparative methods. Therefore, assuming that the majority of the remaining nodes belonging to conserved com-
4.4. Divide algorithm

Figure 4.2: Examples of centers of centered trees (left figure) and of their initial expansion (right figure). Seeds of centers are solid nodes. Dark grey nodes are the rest of centers connected to a seed by solid edges. Light grey nodes are orthologous proteins which are not articulations. Empty nodes are non-orthologous proteins. Dot edges are the rest of edges in the graph. In the second (right) graph dash edges indicate the expansion and connect nodes of centers (zero depth centered trees) with nodes of the first depth centered trees. Nodes on the grey background indicate the overlap among centered trees.

Complexes are neighbors of articulation hubs, we add to each center all its neighbouring orthologous proteins, regardless whether they are or not articulations. We perform this step for all centers in parallel.

We mark these new added proteins with the label of the centers to which they have been added. These new added proteins form the first depth centered trees.

Observe that there may be a non-empty overlap between first depth centered trees (as illustrated in the right part of Figure 4.2).

Parallel Expanding of Trees (Lines 17-27) Successive depths of trees are generated by expanding all nodes with only one label which occur in the last depth of each (actual) centered tree. We add to the corresponding trees all orthologous neighbors of these nodes which are not yet labeled. Then
Figure 4.3: Examples of parallel expansion of trees (left figure) and of the final assigning remaining nodes (right figure). Seeds of centers are solid nodes. Dark grey nodes are the rest of centers connected to a seed by solid edges. Light grey nodes are orthologous proteins which are not articulations. Empty nodes are non-orthologous proteins. Dash edges indicate the process of expansion. Dot edges are the rest of edges in the graph. Nodes on the grey background create the overlap. Numbers are labels of trees assigned to nodes during expansion.

we assign to the newly added nodes the labels of the centered trees they belong to. This process is repeated until it is impossible to add unlabelled orthologous proteins to at least one centered tree.

Observe that each iteration yields to possible overlap between newly created depths (see the left part of Figure 4.3).

**Assigning Remaining Nodes to Trees** (Lines 28-42). The remaining orthologous nodes, that is, those not yet labeled, are processed as follows. First, unlabelled nodes which are neighbors of multi-labeled nodes are added to the corresponding centered trees. Then the newly added nodes are marked with these labels. This process is iterated until there are no unlabelled neighbors of multi-labeled nodes.

Nodes which are not neighbors of any labeled protein are still unlabelled. We assume that they may possibly be part of conserved complexes which
do not contain articulations. Hence we create new sub-trees by joining together all unlabelled orthologous neighbor proteins.

An example of these final steps is shown on the right part of Figure 4.3.

In the end, the algorithm produces the list of subsets of orthologous nodes, where each subset of nodes corresponds to the nodes of one particular tree constructed by the algorithm. The subsets generate induced sub-graphs of the divided PPI network.

**Complexity.** The algorithm divides only orthologs of a given PPI network where the number of all orthologs is $n = |O|$. It performs a parallel breadth-first search (BFS). In general, BFS has $O(|V| + |E|)$ complexity, where $V$ and $E$ denote the number of nodes and edges, respectively. However, the Divide algorithm constructs trees considering only orthologous nodes, so the number of edges, which are traversed, is $|O'|-1$, where $|O'|$ is the number of orthologous vertices of the constructed sub-tree. The possible overlap between trees can increase the number of traversed edges and visited vertices. In the worse case all orthologous vertices are visited by each center (all nodes are in the overlap). So, if the number of centers is $k$, the complexity of Divide is $O(kn)$.

### 4.5 Experimental analysis

The effectiveness and robustness of the proposed divisive method is assessed experimentally in the following two ways.

First, we show that the sub-graphs generated by Divide indeed cover "true" conserved protein complexes. This is done by measuring the overlap of the generated sub-graphs with yeast MIPS curated functional complexes restricted to those proteins belonging to an orthologous pair.

Next, we show that the resulting sub-graphs cover protein complexes computationally predicted by one state-of-the-art alignment algorithm (Koyutürk et al., 2006b), MaWish, in order to investigate whether the sub-graphs contain information that could be used to discover new conserved complexes.

We conduct experiments on the following three pairs of organisms:
Algorithm 1 Divide algorithm

**Input:** $G, G_1$: PPI networks, $O$: orthologous nodes of $G$ with respect to $G_1$

**Output:** $S$: list of subsets of $O$

1: $A = \{\text{orthologous articulations of } G\}$
2: $S = \langle\rangle$
3: repeat {Construction of centers}
4: $\text{root} = \text{element of } A \text{ with highest degree not already occurring in } S$
5: $s = \{\text{root}\} \cup \{\text{neighbors of } \text{root} \text{ in } A \text{ not already occurring in } S\}$
6: $S = \langle s, S \rangle$
7: until all members of $A$ occur in $S$
8: $d = 0$
9: Assign depth $d$ to all elements of $S$
10: Assign label $l_s$ to each $s$ in $S$ and to all its elements
11: for $s$ in $S$ do
12: $s = s \cup \{\text{all neighbors of } s \text{ in } O\}$
13: Assign label $l_s$ to all neighbors of $s$ in $O$
14: end for
15: $d = 1$
16: Assign depth $d$ to all elements of $S$ having yet no depth assigned
17: repeat {Expand one depth centered trees from nodes with one label}
18: $N = \{\text{unlabelled neighbors in } O \text{ of elements in } s \text{ of depth } d \text{ having only one label }\}$
19: for $n$ in $N$ do
20: Assign to $n$ all labels of its neighbors of depth $d$ having only one label
21: for $l_s \in n$ do
22: $s = s \cup \{n\}$
23: end for
24: end for
25: $d = d + 1$
26: Assign depth $d$ to all elements of $S$ having yet no depth assigned
27: until $S$ does not change
28: repeat {Expand one depth centered trees from nodes with multiple labels}
29: $R = \{\text{unlabelled proteins in } O \text{ with at least one multi-labeled protein as neighbor }\}$
30: for $r$ in $R$ do
31: Assign to $r$ all labels of its neighbors
32: for $l_s \in r$ do
33: $s = s \cup \{r\}$
34: end for
35: end for
36: until $S$ does not change
37: repeat {Assign labels to remaining elements}
38: choose an unlabelled element $u$ of $O$
39: $t = \{u\} \cup \{\text{all elements of } O \text{ which can be reached alongside an orthology path from } u\}$
40: Assign label $l_t$ to $t$ and to all its elements
41: $S = \langle t, S \rangle$
42: until $O$ does not contain any unlabelled node
4.6. Divide generates sub-graphs covering "true" protein conserved complexes

- *Saccharomyces cerevisiae* versus *Caenorhabditis elegans* (yeast-worm),
- *Saccharomyces cerevisiae* versus *Drosophila melanogaster* (yeast-fly),
- *Saccharomyces cerevisiae* versus *Homo sapiens* (yeast-human).

Publicly available data were used, available at the web-page of MaWish². These data consist of protein interactions obtained from the BIND (Bader et al., 2001) and DIP (Xenarios et al., 2002) molecular interaction databases, and the list of potential orthologous and paralogous pairs, which are derived using BLAST E-values (for more details see Koyutürk et al., 2006b). Table 4.1 and Table 4.2 report the number of interactions and proteins in the considered species, and the number of potential orthologous pairs between species considered in the alignment task, respectively.

<table>
<thead>
<tr>
<th></th>
<th>S. cerevisiae</th>
<th>C. elegans</th>
<th>D. melanogaster</th>
<th>H. sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>#proteins</td>
<td>5157</td>
<td>3345</td>
<td>8577</td>
<td>4541</td>
</tr>
<tr>
<td>#interactions</td>
<td>18192</td>
<td>5988</td>
<td>28829</td>
<td>7393</td>
</tr>
</tbody>
</table>

Table 4.1: Protein interaction network properties of yeast, worm, fly and human.

<table>
<thead>
<tr>
<th>Pair of species</th>
<th>#orthologous pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae vs C. elegans</td>
<td>2746</td>
</tr>
<tr>
<td>S. cerevisiae vs D. melanogaster</td>
<td>15884</td>
</tr>
<tr>
<td>S. cerevisiae vs H. sapiens</td>
<td>6690</td>
</tr>
</tbody>
</table>

Table 4.2: Number of potential orthologous pairs for considered species: yeast-worm, yeast-fly and yeast-human.

4.6 Divide generates sub-graphs covering "true" protein conserved complexes

Let *Divide* sub-graphs denote the sub-graphs generated by *Divide*. We compared *Divide* sub-graphs with "true" protein conserved complexes. To

²http://compbio.case.edu/koyuturk/software/
this aim, we evaluated the quality of sub-graphs generated by Divide using known yeast complexes catalogued in the MIPS database\(^3\) (Güldener et al., 2005). Category 550, which was obtained from high throughput experiments, is excluded and we retained only manually annotated complexes up to depth 3 in the MIPS tree category structure as standard of truth for quality assessment. From each of these complexes we extracted the subset of proteins consisting of only orthologous proteins, where sets with less than three elements were filtered out. We call the resulting set of proteins *yeast MIPS (conserved) complex*.

Table 4.3 reports the number of yeast sub-graphs and yeast conserved complexes of the alignment tasks for the given pairs of species (yeast-worm, yeast-fly and yeast-human) after and before the application of the filtering procedures above described.

<table>
<thead>
<tr>
<th>Alignment task</th>
<th>#sub-graphs</th>
<th>#yeast MIPS complexes</th>
<th>#yeast MaWish complexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae vs C. elegans</td>
<td>53 (235)</td>
<td>56, 45 (135)</td>
<td>27 (83)</td>
</tr>
<tr>
<td>S. cerevisiae vs D. melanogaster</td>
<td>119 (408)</td>
<td>111, 99 (205)</td>
<td>99 (411)</td>
</tr>
<tr>
<td>S. cerevisiae vs H. sapiens</td>
<td>67 (253)</td>
<td>77, 63 (161)</td>
<td>57 (276)</td>
</tr>
</tbody>
</table>

Table 4.3: Number of yeast sub-graphs and yeast conserved complexes for a given alignment task: yeast-worm, yeast-fly or yeast-human. In brackets the number of sub-graphs and complexes before removing sets with less than three elements is given. The second number in the yeast MIPS complexes column is the number of complexes after big-sized complexes have been removed.

The *intersection rate* between a sub-graph and a complex is used, computed as follow. Let \( G = (U, E) \) be a sub-graph and let \( C \) be a protein complex of one organism. The intersection rate of \( G \) and \( C \) is

\[
\frac{|U \cap C|}{|C|}.
\]

In case more *Divide* sub-graphs have equal intersection rate with a given complex, we chose the sub-graph of smallest size. This sub-graph provides

\(^3\)http://mips.helmholtz-muenchen.de/genre/proj/yeast/
4.6. Divide generates sub-graphs covering "true" protein conserved complexes

a best coverage of the considered complex, because it needs the smallest number of proteins to achieve that intersection rate.

The relation between the intersection rate of yeast Divide sub-graphs and a "true" complex, and the size of a "true" complex are shown in the left column of Figure 4.4 for yeast-worm, yeast-fly or yeast-human alignment task. Low intersection rates mostly correspond to complexes of big size (see left upper part of the plots).

Because conserved complexes have in general small size, we incorporated this prior information in our analysis and filtered out complexes of big size from the list of yeast MIPS complexes, since they were not considered to be conserved. To this end, we used the conserved complexes predicted by MaWish (see also the next section). For yeast-worm and yeast-human the biggest yeast MaWish complex has size 12, for the yeast-fly alignment task the biggest MaWish complex consists of 21 proteins. Using these parameter values for the threshold to filter out yeast MIPS complexes considered too large, we got 45 yeast MIPS complexes w.r.t. worm, 99 complexes w.r.t. fly and 63 complexes w.r.t. human (see Table 4.3). As shown in Table 4.4 the average intersection rate increased for the small yeast MIPS complexes while the number of considered complexes does not decrease significantly.

Another issue concerns the selection of only one Divide sub-graph when computing the intersection rate with a complex. Divide sub-graphs having equal intersection rate with a complex may cover that complex in different ways. Therefore, one should consider the contribution to the coverage of that complex provided by all these sub-graphs. This may be formalized by defining a so-called union intersection rate as follows. Let \( S \) be a set of Divide sub-graphs having the same intersection rate with a complex \( C \). The union intersection rate is

\[
\left| \bigcup_{G(U,E) \in S} U \cap C \right| / \left| C \right|
\]

The average union intersection rate between yeast MIPS complexes and sub-graphs is shown in Table 4.4 for three alignment tasks. The union intersection rate is higher than the intersection rate. Highest values are obtained for small-size complexes. For each alignment task, more than 70% coverage
Figure 4.4: Intersection rate vs. size of yeast complexes for the alignment task. Left column: yeast MIPS complexes. Right column: yeast MaWish complexes.
of yeast MIPS complexes is achieved. This means that some yeast MIPS conserved complexes are split among sub-graphs, hence different parts of conserved complexes can be discovered by searching in these sub-graphs.

<table>
<thead>
<tr>
<th>Considered yeast complexes</th>
<th>C. elegans</th>
<th>D. melanogaster</th>
<th>H. sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>small-sized (union)</td>
<td>71.4%</td>
<td>75.7%</td>
<td>81.9%</td>
</tr>
<tr>
<td>all (union)</td>
<td>64.5%</td>
<td>71.3%</td>
<td>74.8%</td>
</tr>
<tr>
<td>small-sized</td>
<td>64.0%</td>
<td>68.0%</td>
<td>69.3%</td>
</tr>
<tr>
<td>all</td>
<td>56.0%</td>
<td>64.2%</td>
<td>63.4%</td>
</tr>
</tbody>
</table>

Table 4.4: Average of (union) intersection rate of yeast MIPS complexes and sub-graphs given an alignment task: yeast-worm, yeast-fly or yeast-human.

These results indicate that Divide is able to generate sub-graphs that highly cover ”true” conserved complexes. Lower intersection rate for yeast MIPS complexes could be due to the fact that functional complexes in MIPS database are not biased on protein interaction conservation across species. Nevertheless, we achieved a satisfactory intersection rate for small-sized complexes, which are more likely to be (part of) conserved protein complexes.

### 4.7 Comparison of Divide sub-graphs with predicted conserved complexes

Here we investigate how Divide constrains the search process of MaWish, and whether the sub-graphs generated by Divide cover those complexes produced by MaWish. To this end, we used the conserved complexes predicted by this alignment method. In general, conserved complexes computed by network alignment methods tend to highly overlap with each other, that means they share many proteins in common. Therefore, we processed the MaWish conserved complexes by the clique-rule merging procedure as described in Jancura and Marchiori (2010, see also Section 5.3.3) to reduce the number of redundant solutions. Finally, the complexes consisting of one or two proteins were filtered out. We call the resulting sets MaWish
complexes.

In the right column of Figure 4.4 one can observe that a number of yeast MaWish complexes are fully covered and many of those, which are not fully covered, intersect with a sub-graph at a rate higher than 0.5.

Next we computed the average intersection rate of MaWish complexes for each of the considered alignment tasks of yeast-worm, yeast-fly and yeast-human. For a given pair of organisms, we computed the number of conserved complexes for the first and for the second organism, and the intersection rate between the complexes and sub-graphs of the first organism and of the second organism, respectively. In all cases, we got almost or more than 80% coverage of conserved complexes (see Table 4.5).

<table>
<thead>
<tr>
<th>Alignment task</th>
<th>#conserved MaWish complexes</th>
<th>intersection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae vs C. elegans</td>
<td>27, 24</td>
<td>87.0, 91.7</td>
</tr>
<tr>
<td>S. cerevisiae vs D. melanogaster</td>
<td>99, 80</td>
<td>79.9, 84.8</td>
</tr>
<tr>
<td>S. cerevisiae vs H. sapiens</td>
<td>57, 63</td>
<td>84.7, 89.8</td>
</tr>
</tbody>
</table>

Table 4.5: Average intersection rate of MaWish conserved complexes and sub-graphs for a given alignment task. In each column, the first number contains the number of conserved MaWish complexes of yeast and the second one the number of conserved complexes of the second organism in the considered alignment task.

4.8 Conclusions

In this chapter we introduced a heuristic algorithm, Divide, for dividing protein interaction networks in such a way that conserved functional complexes are covered by generated sub-graphs. To the best of our knowledge, this is the first algorithm for this task. The possible application of this algorithm is to narrow network search space in order to enhance or to allow more efficient mining of PPI networks for evolutionary conserved complexes.

The selection of centers is biased on the orthology information but it can be changed for another property. Hence, the Divide algorithm can
be applied on other type of networks where restricting the search space is desirable and there is no need to analyse networks globally.

We showed experimentally that the sub-graphs that were generated by Divide covered part of predicted conserved complexes. In some cases these sub-graphs covered different parts of one conserved complex.

Last but not least, another advantage of applying the Divide algorithm is that it allows one to parallelize computational methods in order to detect conserved complexes. For instance, a protein network alignment method can be run independently on each possible pair of constructed PPI sub-graphs generated by Divide.
Chapter 5

Modular alignment of protein interaction networks

The increasing growth of data on PPI networks has boosted research on their comparative analysis. In particular, recent studies proposed various models and algorithms for performing network alignment, that is, the comparison of networks across species for discovering conserved functional complexes. Here, we apply the Divide algorithm to divide PPI networks prior to their comparison in order to perform the network alignment in a modular fashion, by acting on pairs of resulting small sub-graphs from different species. The thorough comparative analysis proves the ability of the proposed modular approach to enhance the performance of state-of-the-art alignment methods.\(^1\)

\(^1\)This chapter is based on the following works:

5.1 Introduction

Comparing PPI networks of distinct species is one of the recent approaches in computational biology in order to identify evolutionary common protein complexes and pathways, or to determine functional orthologous proteins (Sharan and Ideker, 2006). As consequence, these methods, known as protein network alignment methods, have a great potential for inference of protein interactions (Dutkowski and Tiuryn, 2009, Schwartz et al., 2009), protein function prediction (Kolar et al., 2008, Sharan et al., 2005b, Singh et al., 2008b), or, in general, for bringing new insights into understanding evolutionary processes and evolution as such (Erten et al., 2009, Tan et al., 2007, Yosef et al., 2009).

Protein network alignment algorithms typically model this problem by means of a merged graph representation of the networks to be compared, called alignment (or orthology) graph, and then formalize the problem of searching (merged) conserved complexes or functional orthologs in the alignment graph as an optimization problem. Due to the computational intractability of the resulting optimization problem, greedy algorithms are commonly used.

An alternative approach to avoid the computational burden imposed by these models is to restrict an alignment algorithm to be performed only on local regions of PPI networks much smaller than the entire networks considered for the comparison. This strategy, called modular network alignment (Jancura et al., 2008a), may be especially efficient for detection of conserved functional protein complexes as those have relatively small size (at most tenths of proteins) in comparison with the total network size (hundreds or thousands of proteins). Moreover, conserved complexes are local regions of compared PPI networks which share high topological and genomic similarity.

Here, we present an application of modular network alignment using the Divide algorithm (Jancura et al., 2008b, see also Section 4.4). The Divide method is designed to identify sub-networks consisting of orthologous proteins of a given PPI network such that the sub-networks contain potential conserved complexes and we test its ability to be used as a pre-processing
step prior applying state-of-the-art alignment methods for discovering evolutionary conserved protein modules.

To this end, we consider two case studies of modular network alignment with Divide. In the first case study, Divide is used to generate sub-graphs, which are then pairwise merged using the network merging model of MaWish (Koyutürk et al., 2006b). We apply iterative exact search to the resulting alignment graphs. Results of experiments show ability to detect a high number of accurate conserved complexes. In the second case study, Divide is used for enhancing an existing method for discovering conserved functional complexes, called MNAligner (Li et al., 2007). MNAligner consists of two main steps: first, candidate functional complexes within one species are detected using a clustering algorithm (MCODE); next, an exact optimization algorithm is applied for matching the resulting candidate functional complexes with the PPI network of the other species in order to extract conserved complexes. Results of experiments show that by applying Divide to orthologs nodes prior to clustering enhances the performance of this algorithm.

This chapter is organized as follows. First, we discuss related works. Next, in Section 5.3 we describe two instances of modular network alignment with Divide. Section 5.4 defines evaluation criteria for the quality assessment of detected solutions. Then, in Section 5.5 we perform modular network alignments and discuss their outcomes. Finally, we present our conclusions in Section 5.6.

5.2 Related work

Recent overviews of approaches and issues in comparative biological networks analysis have been presented by Sharan and Ideker (2006), Srinivasan et al. (2007) and Jancura and Marchiori (2012) since the first formulation of network alignment introduced by Kelley et al. (2003).

In general, network alignment methods have been proposed for discovering conserved metabolic pathways, conserved functional complexes, and for detecting functional orthologs. For instance, in Kelley et al. (2003) in-
Chapter 5. Modular network alignment

Introduced an approach for detecting conserved metabolic pathways between two species. A local protein network alignment method based on this approach was proposed to discover conserved complexes (Sharan et al., 2004, 2005a). This method was further extended to the alignment of multiple species by Sharan et al. (2005b). Moreover, the approach of Kelley et al. (2003) motivated Bandyopadhyay et al. (2006) to develop a method for identifying functional orthologs.

One can identify two main network alignment categories. Local network alignment, that identifies the best local mapping for each local region of similarity between input networks, and global network alignment, that searches for the best single mapping across all parts of the input networks, even if it is locally sub-optimal in some regions of the networks. If a method aligns networks of just two species, it is called pairwise network alignment, while if it can handle more than two networks, it is called multiple network alignment.

The main goal of local protein network alignment is to detect conserved protein complexes across species, by searching for local regions of input networks having both high topological similarity between the regions and high sequence similarity between proteins of these regions. Many pairwise local network alignment techniques have been introduced in recent years (see, e.g. Hirsh and Sharan, 2007, Koyutürk et al., 2005, Liang et al., 2006, Narayanan and Karp, 2007, Sharan et al., 2005a, Tian and Samatova, 2009). In particular, Berg and Lässig (2006) introduced an alignment framework based on Bayesian theory. Other approaches embed additional information into the local protein network alignment task (Ali and Deane, 2009, Guo and Hartemink, 2009).

A first attempt to perform multiple network alignment using three species was done by Sharan et al. (2005b). However, the method scales exponentially with the number of input species. Thus, new methods for aligning multiple species have been proposed (Dutkowski and Tiuryn, 2007, Flannick et al., 2006, 2009, Kalaev et al., 2009).

The main goal of global protein network alignment is functional orthologs detection, because, in contrast to local network alignment, each node in an input network is either matched to one node in the other net-
work or has no match in the other network (Singh et al., 2007). Global protein network alignment, however, can be also used for identifying conserved complexes.

The first systematic identification of functional orthologs based on protein network comparison was done by Bandyopadhyay et al. (2006). Singh et al. (2008b) explicitly used global multiple network alignment for detecting functional orthologs.

The first formal global network alignment method was introduced by Singh et al. (2007). This method has been followed by more works on global pairwise network alignment (Chindelevitch et al., 2010, Evans et al., 2008, Klau, 2009, Zaslavskiy et al., 2009). Flannick et al. (2009), Liao et al. (2009) and Singh et al. (2008a) tackled global alignment of multiple species.

While the above works focus on alignment of entire networks, we deal with protein networks pre-processing prior to their alignment, in order to perform modular network alignment (Jancura et al., 2008a,b).

In modular network alignment only certain regions of the compared networks are aligned. For instance Gerke et al. (2007) first detect hubs and protein clusters independently in both PPI networks considered for alignment. Then the method performs sequence comparison only between proteins of those structures identified in each network where the proteins around hubs may still be exploited. Subsequently, network topologies on selected pairs of these local regions are compared in order to extract evolutionary conserved network structures. The method tries to minimize the number of protein sequence comparisons to be performed prior the network alignment. As consequence, the method does not use putative orthologous relationship as prior information but rather it identifies them during construction of the local network regions to be aligned. However, this restricts the procedure to be directly applicable for alignment models other than the one proposed by Gerke et al. (2007).

More general approach was used by Li et al. (2007) who applied their pairwise global network alignment algorithm to identify also conserved protein modules between species. Specifically, Li et al. (2007) at first extract putative protein complexes from one of the PPI networks by applying a state-of-the-art network clustering on the entire network. Then these com-
plexes are consequently aligned with the other PPI network using the proposed global alignment method in order to discover their conserved parts.

Similarly, Erten et al. (2009) and Yosef et al. (2009) also opted for modular network alignment to study the evolution of protein functional modules in PPI networks. Much like Li et al. (2007) they applied clustering techniques independently on each investigated PPI network to detected putative functional complexes of every species under the study. Thereafter only the complexes were aligned across species.

In this work, we use the Divide algorithm (Jancura et al., 2008b) to pre-process PPI network prior network alignment. The main difference between Divide and typical network clustering methods is Divide constructs sub-networks of a PPI network containing possibly multiple conserved complexes, which still need to be extracted, while clustering techniques usually detect ad-hoc protein complexes. The secondary difference is Divide runs only on the sub-network of the entire PPI network consisting of putative orthologous proteins, where the above-mentioned studies on modular alignments used clustering methods without exploiting the orthologous information available in PPI networks. Although, one can obviously apply clustering only on specific sub-networks of interest.

5.3 Modular network alignment with Divide

In this section we illustrate different applications of Divide in order to enhance the performance of state-of-the-art alignment methods. Specifically, we describe two instances of modular alignment approach employing Divide that use distinct protein network comparative methods.

In the first case, we consider an instance of modular local network alignment, called DivAfull (Jancura et al., 2008a). DivAfull employs Divide to generate sub-graphs, the MaWish alignment model (Koyutürk et al., 2006b) to align them, and iterative exact search to detect all possible solutions from the generated alignments. Therefore, application of Divide allows one to improve the search process by replacing the greedy search procedure of MaWish with an exact search algorithm.
5.3. Modular network alignment with Divide

In the second case we consider an instance of modular global network alignment. Specifically, we show how already the modular approach of Li et al. (2007), which uses the global alignment method MNAAligner (Li et al., 2007) can be enhanced by prior application of Divide. In order to detect conserved complexes using MNAAligner a clustering algorithm is applied which detects potential protein complexes in one PPI network. The resulting complexes are then aligned with the second PPI network and the conserved protein (sub-)complexes are detected. Here, we apply Divide before clustering in order to bias the search for complexes towards regions centred around articulation hubs. Results of experiments indicate that this is an effective way of enhancing the discovery of conserved complexes using MNAAligner.

We describe DivAfull and MNAAligner more in detail in the next two sections. Then we discuss a post-processing step in order to reduce redundancy among alignment solutions.

5.3.1 DivAfull

DivAfull uses the Divide procedure to generate sub-graphs for each of PPI networks given by species to be compared. Next, pairs of the sub-graphs from different species are merged using the MaWish network alignment model.

In that model, a weighted alignment graph is constructed from a pair of PPI networks and a similarity score $S$, which quantifies the likelihood that two proteins are orthologous, is computed. A node in the alignment graph is a pair of orthologous proteins. Each edge in the alignment graph is assigned a weight that is the sum of three scoring terms: for protein duplication, mismatches for possible divergence in function, and match of a conserved pair of orthologous interactions. We refer to Koyutürk et al. (2006b) for a formal description of these terms.

Induced sub-graphs of the resulting weighted alignment graph with total weight greater than a given threshold are considered as relevant alignments. Each relevant alignment corresponds to two putative conserved complexes, one for each species.
Chapter 5. Modular network alignment

After merging the two PPI networks we search for these sub-graphs. This problem is reduced to the (optimization) problem of finding a maximal induced sub-graph. To tackle this problem, the search part of MaWish consists of an approximation greedy algorithm based on local search, because the maximum induced sub-graph problem is NP-complete. This greedy algorithm selects at first one seed which can likely contribute at most to the overall weight of a potential sub-graph. The seed is expanded by adding (removing) nodes to (from) the sub-graph while the actual sub-graph weight increases.

In contrast, DivAfull applies iteratively an exact optimization algorithm (Wolsey, 1998) for searching relevant alignments (maximum weighted induced sub-graphs) in the alignment graphs produced by merging possible pairs of the Divide sub-graphs, since the constructed alignment graphs are small in size. Two Divide sub-graphs form a possible pair for merging if there exists a putative orthologous pair of proteins between them. The iterative search algorithm is described in detail below.

Search Algorithm

First, an exact optimization algorithm for finding the maximum weighted induced sub-graph is applied on an alignment graph. Then the process is iterated by adding at each iteration the constraint which bounds the weight of the induced sub-graph by the weight of the solution found in the previous iteration.

Formally, let $f$ be a function which computes the weight of a sub-graph in an input graph and $C$ be a set of constraints which defines an induced sub-graph of the input graph. Then we want to maximize the function $f$ on the set defined by constrains $C$, that is, to solve the following optimization problem:

$$
\text{opt} = \max_{C} f
$$  \hfill (OptP)

Algorithm 2 illustrates the resulting full-search procedure which uses the above constrained optimization problem at each iteration with different bound on the maximum allowed weight.
Algorithm 2 Full Search Algorithm

Input: $G$: alignment (sub-)graph, $\epsilon \geq 0$
Output: List of heavy induced sub-graphs of $G$ with weight $> \epsilon$
1: Formulate the problem of maximum induced sub-graph for $G$ as ($OptP$)
2: $maxweight = \infty$
3: $C = C + \{opt < maxweight\}$
4: while $maxweight > \epsilon$ do
5: solve ($OptP$) by an exact method
6: if $opt > \epsilon$ then
7: record discovered solution
8: end if
9: $maxweight = opt$
10: end while

5.3.2 Divide and MNAligner

MNAligner is a general tool for global alignment of molecular networks. It formalizes the problem of finding an optimal mapping between similar nodes of two different networks as an integer quadratic programming optimization problem which is relaxed to quadratic optimization problem (QP). The optimal integer solution is ensured if appropriate sufficient conditions on the objective function are satisfied. However, QP may have an integer solution also if the conditions are not satisfied. We refer to Li et al. (2007) for the detailed description of the objective function and alignment algorithm.

Direct application of MNAligner is feasible only when small PPI networks are considered. Furthermore, the alignment algorithm finds the global mapping but does not search for structures of interest in this mapping, such as being dense sub-graphs. Therefore an additional search for such structure is needed either before or after the alignment. For instance, in one of the applications of MNAligner described in Li et al. (2007) two large PPI networks are aligned in order to detect conserved complexes. To overcome the problem of large network size, and to bias the search towards detection of protein complexes, the clustering algorithm MCODE (Bader and Hogue, 2003) is applied to one of the networks prior the alignment. This algorithm generates a set of clusters representing potential functional protein complexes. Each of these clusters is aligned with the second PPI network, resulting in the detection of two conserved sub-networks, one for
each species being compared.

Application of Divide yields sub-graphs representing regions of interest which potentially cover a number of conserved complexes. We test whether the application of Divide prior to the use of MCODE and MNAligner enhances the discovery of conserved complexes. Specifically, given two PPI networks \( G_1 \) and \( G_2 \), we divide \( G_1 \) using Divide. Each of the resulting sub-graphs is further processed by MCODE and aligned with \( G_2 \) using MNAligner. In this way two collections of conserved complexes are generated, one for each species. We repeat this process by dividing \( G_2 \) instead of \( G_1 \). This gives again two collections of complexes. For each species, the union of the collections of detected complexes from that species is considered. In such way we get two complete collections of possible conserved complexes detectable by MNAligner. As the complexes of each final collections after the union may considerably overlap and be highly redundant we do a post-processing of each final collection as described below in Section 5.3.3.

These results are afterwards compared with complexes produced when MCODE and MNAligner are directly applied to the PPI networks induced by orthologs. Specifically, we cluster the (orthologous sub-network of) \( G_1 \) and align the resulting clusters with \( G_2 \). We repeat the process by interchanging the role of \( G_1 \) and \( G_2 \). We again do the union of detected complexes for each species and perform the same post-processing method on final collections (see below Section 5.3.3).

In this way, we allow a fair comparison of results when only MCODE and MNAligner are applied and when Divide is introduced prior these steps. We restrict the use of MCODE on sub-networks induced by orthologs, because Divide divides only the orthologs of a PPI network.

5.3.3 Handling redundant alignments and complexes

A general issue in network alignment methods is that the solutions produced usually considerably overlap with each other; in other words they are highly redundant. Specifically, two clusters of nodes are said to be redundant if more than \( r \% \) of the nodes in the smaller complex occur in the other
complex, where $r$ is a threshold value that determines the extent of allowed overlap between clusters.

Recall that most network alignment methods construct an alignment graph, which is a merged representation of the protein interaction networks being compared. Then, alignment solutions or relevant alignments are network structures of interest found by searching the alignment graph. Each discovered alignment corresponds to a set of complexes, one for each given organism, which are conserved to each other. Thus, a set of alignment solutions gives separate collections of conserved complexes for the species being compared.

Obviously, one may observe the redundancy at two levels: alignment level and protein level. The first level is when alignments found in the alignment graph highly overlap. The second one is when conserved protein complexes in one collection highly overlap.

As mentioned above, in the experimental analysis of this study we use two alignment methods, MaWish and MNAligner. At the alignment level, MaWish filters out redundant solutions ($r = 80\%$) retaining only alignments with bigger score. MNAligner is a global network alignment method where the computed mapping between orthologs is a one-to-one mapping resulting in solutions that do not overlap.

At the protein level MaWish does not handle possible occurrence of redundant complexes. Moreover, despite the fact that MNAligner performs global alignment, our application of MNAligner may produce intersecting complexes, because for each species the final collection of conserved complexes is the union of collections previously produced by two consecutive alignments by inter-changing the role of the networks being aligned. In addition, the sub-graphs generated by Divide can overlap and hence the complexes extracted from them may overlap too. Therefore, in both instances of modular network alignment here considered, we will have to handle redundant protein complexes.

In general if two complexes have a high intersection, one of them is discarded (see, e.g. Supporting Methods of Sharan et al., 2005b). However, this approach for handling redundancy is not very satisfactory, since detected conserved complexes could possibly cover part of a ‘true’ functional
module either due to constraints on the topology and homology similarity, or due to missing interactome data. Therefore, detected complexes having high overlap may still represent different parts of one bigger module.

In Liang et al. (2006) the following alternative method is proposed for merging redundant solutions. If two clusters are highly intersecting then they are merged into a single cluster by taking the union of the two clusters. Three or more clusters are merged by the rule of single linkage, that is, the merging relation is transitive. We refer to this method as chain-rule merging. A drawback of this procedure is that it may merge protein complexes whose intersection is not any more above the required threshold due to the transitive relation used. Therefore parts of different modules might be merged. Furthermore, application of the chain-rule merging can produce one or few very big modules containing several possible functional complexes.

These observations motivate the introduction of the following procedure for dealing with highly intersecting complexes. Specifically, we modify the chain-rule merging as follows. A set of complexes is merged if every possible pair of complexes contained in this set is redundant.

If we represent complexes by means of nodes and connect two nodes by an edge if they are redundant then the problem of finding a maximal set of complexes which can be merged according to the above rule can be reduced to the problem of finding a maximal clique in that graph. Consequently, finding all such maximal sets is equivalent to the problem of finding all maximal cliques, which is an intractable optimization problem. Nevertheless, in our setting the resulting graph is rather sparse and contains relatively few nodes, which allows us to apply an exact algorithm for finding all maximal cliques in graph (here we use the algorithm of Bron and Kerbosch (1973)).

We refer to the modified merging procedure as clique-rule merging. In our experimental analysis the redundancy threshold \( r = 80\% \) is used.

As the final step, after applying clique-rule merging on a given collection of conserved complexes, we retain only complexes of size greater than or equal to 3 proteins.
5.4 Evaluation criteria for conserved complexes

We assess the performance of alignment methods by measuring the quality of detected complexes. A functional module may perform one or more functions in an organism and all proteins contained in that module are associated with these functions. Based on this assumption, computationally derived protein complexes may serve for predicting function of proteins. Then the quality of a complex can be assessed by the function prediction of the proteins it contains.

Therefore, we measure the enrichment of functional annotations of the protein set in a complex, as entailed by the gene ontology (GO)\textsuperscript{2} annotation (Ashburner et al., 2000), using one of the well-established tools, the Ontologizer\textsuperscript{3} (Bauer et al., 2008). Ontologizer measures statistical significance of an enrichment and assigns to the complex a p-value for each enriched function. The p-value is corrected for multiple testing by a classic Bonferroni correction procedure. Furthermore, Ontologizer also constructs a hierarchical directed acyclic graph (DAG) consisting of all significantly enriched annotations and all their ancestor annotations up to the root in the whole GO hierarchy. Given a DAG of enrichments, the level of an annotation is equal to the length of the shortest path from the root of GO hierarchy present in the DAG to that annotation.

A complex can be used as protein function predictor if the following criteria are satisfied:

1. a certain GO annotation is significantly enriched by the proteins in the complex (p-value < 0.05);

2. at least half of the proteins in the complex has this significant annotation;

3. the annotation is at least at GO level four from the root in GO hierarchy.

\textsuperscript{2}http://www.geneontology.org/
\textsuperscript{3}http://compbio.charite.de/index.php/ontologizer2.html

103
Chapter 5. Modular network alignment

In such a case the significantly enriched GO annotation of the complex is used to predict protein function of each of the proteins in that complex. If a complex does not satisfy the above conditions, no prediction can be made. Similar criteria were used by, e.g. Liang et al. (2006). The condition on GO hierarchy guarantees that the prediction about biological functions is sufficiently specific and informative (Yon Rhee et al., 2008).

We validate the accuracy of the predictions, and consequently the quality of a protein complex, in a way similar to that proposed by Deng et al. (2003). Specifically, given a protein complex and the corresponding DAG of enrichments, we restrict our validation only to the annotations which are present in the DAG and are at GO level four or higher. A protein $p$ of the complex having such annotations is assumed to be not annotated and its functions are predicted. The predictions are then compared with the annotations of the protein $p$. The method is repeated for all annotated proteins in the cluster. In the end, for each protein $p$ we have:

- $A_p$: the number of annotated functions for the protein $p$.
- $P_p$: the number of predicted functions for the protein $p$.
- $O_p$: the size of the overlap between the set of annotated functions and the set of predicted functions for the protein $p$.

Given this scheme, precision (PR) and recall (RC) are computed for each complex $C$ as follows:

$$PR(C) = \frac{\sum_{\forall p \in C} O_p}{\sum_{\forall p \in C} P_p},$$

$$RC(C) = \frac{\sum_{\forall p \in C} O_p}{\sum_{\forall p \in C} A_p}.$$

In the case of no prediction, precision and recall are set to zero. When both precision and recall are close to one then function prediction of a protein complex is good. Therefore, we also use the following well-established
5.4. Evaluation criteria for conserved complexes

measure in information retrieval (Rijsbergen, 1979) as suggested by Handl et al. (2005), the $F$-measure (FM), defined as

$$FM(C) = \frac{2 \cdot PR(C) \cdot RC(C)}{PR(C) + RC(C)},$$

where we assume that both precision and recall are equally important. We use the above evaluation measures to validate the quality of a predicted complex with respect to its ability to model the functions of the proteins it contains.

In order to assess whether Divide leads to the discovery of conserved complexes having a new putative function, we introduce the following two additional measures, functional ratio ($FNR$) and coverage ratio ($CVR$).

Let $A$ be the collection of all functions predicted by the complexes detected by the original method and let $B$ be the collection of all function predicted by the complexes detected by the combined method. Furthermore, denote by $C_X$ the set of all complexes which are predicted to have a function from the function collection $X$. Then

$$FNR = \frac{|B \setminus A|}{|B|},$$

$$CVR = \frac{|C_B \setminus A|}{|C_B|}.$$

The first measure, $FNR$, computes the ratio of new functions discovered over the set of all functions discovered by the combined method. The latter one computes the ratio between the number of complexes which are predicted to have the new functions and the total number of complexes detected by the combined method.

Notice that all measures above defined treat each species separately rather than explicitly evaluating the conservation hypothesis implied by each pair of conserved complexes aligned. Such evaluation could, in principle, be performed by comparing the results to a reference set of conserved modules (Yosef et al., 2008). To date, however, most such references are not comprehensive enough and contain only a small number of cases to
learn from (Yosef et al., 2008). One exception is the Biocarta\textsuperscript{4} (Nishimura, 2001) database which contains many human-mouse conserved pathways.

Finally, it should be also noted that the functional annotations for the annotated proteins are incomplete. Thus, we may have a high confidence in the assignment of the function to a protein based on the GO annotation. However, that protein can have a particular true function which has not yet been annotated, that is, it has not been experimentally validated.

5.5 Results

We run the proposed modular alignment approaches on PPI networks of the two following species: \textit{Saccharomyces cerevisiae} (yeast) and \textit{Caenorhabditis elegans} (worm). As these species are well studied, they are typically used for performance assessment of network alignment methods (see e.g. Koyutürk et al., 2006b, Sharan et al., 2005b). We use publicly available data at the webpage of the MaWish method\textsuperscript{5}, which compiles interactions from BIND (Bader et al., 2001) and DIP (Xenarios et al., 2002) molecular interaction databases. The yeast PPI network consists of 5157 proteins and 18192 interactions, and the worm PPI network consists of 3345 proteins and 5988 interactions. Moreover, the data already contain the list of potential orthologous pairs, which are derived using BLAST E-values (for more details see Koyutürk et al., 2006b). 2746 potential orthologous pairs created by 792 proteins in \textit{S. cerevisiae} and 633 proteins in \textit{C. elegans} are identified.

We present the results as follows. We summarize the application of \textit{Divide} algorithm on particular PPI networks. Then we show how the iterative exact search of \textit{DivAfull} improves on \textit{MaWish} results. Finally, we discuss results of \textit{MNAligner} combined with \textit{Divide}.

\begin{footnotesize}
\textsuperscript{4}http://www.biocarta.com/genes/allPathways.asp
\textsuperscript{5}http://compbio.case.edu/koyuturk/software/
\end{footnotesize}
5.5. Results

5.5.1 Application of Divide

Results of application of the Divide algorithm to the PPI networks of Saccharomyces cerevisiae and Caenorhabditis elegans are following.

For Saccharomyces cerevisiae, 697 articulations, of which 151 orthologs, were computed, and 83 centers were constructed from them. Expansion of these centers into centered trees resulted in 639 covered orthologs. The algorithm assigned the remaining 153 orthologous proteins to 152 new sub-trees.

For Caenorhabditis elegans, 586 articulations, of which 158 orthologs, were computed, and 112 centers were constructed from them. Expansion of these centers into centered trees resulted in 339 covered orthologs. The algorithm assigned the remaining orthologous 294 proteins to 288 new sub-trees.

We observed that the last remaining orthologs assigned to sub-trees were 'isolated' nodes, in the sense that they were rather distant from each other and not reachable from ortholog paths stemming from centers.

We obtained 235 sub-trees for Saccharomyces cerevisiae and 400 sub-trees of Caenorhabditis elegans. Nodes of each such tree induce a PPI sub-graph.

5.5.2 DivAfull and MaWish

DivAfull constructs alignment graphs between each two PPI sub-graphs containing more than one orthologous pair. In such way, we obtained 884 alignment graphs, where the biggest one consisted of only 31 nodes.

We applied Algorithm 2 to each of the resulting alignment graphs. Zero weight threshold ($\varepsilon = 0$) was used for considering an induced sub-graph as a heavy sub-graph or a legal alignment. Redundant graphs were filtered using $r = 80\%$ as the threshold for redundancy.

DivAfull discovered 151 solutions (relevant alignments) while MaWish yielded 83 solutions. Between these two set of solutions we found 70 redundant alignments, whose pair of weights are plotted on the left part of Figure 5.1. Among these, 48 (31.8\% of DivAfull results) were equal (crosses in
the diagonal) and 22 (14.6%) different. 8 (5.3%) (diamonds below the diagonal) with better DivAfull alignment weight, and 13 (8.6%) (circles above the diagonal) with better MaWish alignment weight (for 1 (0.7%) pair it was undecidable because of rounding errors during computation).

DivAfull found 81 (53.6%) new alignments, that is, not discovered by MaWish. The right plot of Figure 5.1 shows the binned distribution of weights of these alignments, together with the new 17 ones discovered by MaWish but not by DivAfull. There is no significant difference between the overall weight average of the DivAfull (0.8) and the MaWish (0.86) results.

Figure 5.1: Analysis of all alignments discovered by MaWish and DivAfull. Left figure: Distribution of pairs of weights for paired redundant alignments, one obtained from MaWish and one from DivAfull. Weights of alignments found by DivAfull are on the x-axis, those found by MaWish on the y-axis. '+' is a paired redundant alignment. Right figure: Interval weight distributions of non-redundant alignments discovered by MaWish and DivAfull. The x-axis shows weight intervals, the y-axis the number of alignments in each interval.

Further, we investigate conserved complexes derived from the alignments discovered. Recall, each set of discovered alignments gives two collections of conserved complexes, one for each species being compared, which are processed by clique-rule merging algorithm and only complexes of size
greater than 2 are considered.

DivAfull discovered a higher number of protein complexes than MaWish and the same is observed when only those complexes which satisfy the criteria for being a functional predictor are considered. Specifically, for *Saccharomyces cerevisiae* DivAfull found 46 complexes of which 39 are functional predictors, and for *Caenorhabditis elegans* DivAfull found 28 complexes of which 18 are functional predictors. In contrast, MaWish found 27 complexes of which 24 are potential predictors for *Saccharomyces cerevisiae* and 24 complexes of which 13 are functional predictors for *Caenorhabditis elegans*.

We measured the GO enrichment of these complexes and computed the average of their precisions, the average of their recalls, and the average of their F-measures. The results are reported in Table 5.1 and Table 5.2 for *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, respectively.

For *Saccharomyces cerevisiae*, when considering all modules, we observe lower average precision and average F-measure of DivAfull modules than of MaWish complexes (the upper part of Table 5.1). However, the difference in F-measures is subtle and average recalls are same. Thus, complexes of both methods are, in total, of comparable quality. When focused on functional predictors (the bottom part of Table 5.1), DivAfull clearly outperforms MaWish functional predictors.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Modules</th>
<th>Precision (±δ)</th>
<th>Recall (±δ)</th>
<th>F-measure (±δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DivAfull</td>
<td>46</td>
<td>0.73 (±0.33)</td>
<td><strong>0.52</strong> (±0.30)</td>
<td>0.59 (±0.31)</td>
</tr>
<tr>
<td>MaWish</td>
<td>27</td>
<td><strong>0.75</strong> (±0.30)</td>
<td><strong>0.52</strong> (±0.26)</td>
<td><strong>0.60</strong> (±0.28)</td>
</tr>
<tr>
<td>DivAfull</td>
<td>39</td>
<td><strong>0.86</strong> (±0.13)</td>
<td><strong>0.61</strong> (±0.22)</td>
<td><strong>0.70</strong> (±0.20)</td>
</tr>
<tr>
<td>MaWish</td>
<td>24</td>
<td>0.84 (±0.13)</td>
<td>0.58 (±0.20)</td>
<td>0.67 (±0.18)</td>
</tr>
</tbody>
</table>

Table 5.1: The average of precisions, the average of recalls, and the average of F-measures of yeast protein modules. The upper part reports results for all complexes, the bottom part for all functional predictors.

For *Caenorhabditis elegans*, when considering all modules, a better average functional enrichment is achieved for DivAfull modules (the upper part of Table 5.2). Considering all functional predictors, MaWish complexes
Chapter 5. Modular network alignment

<table>
<thead>
<tr>
<th>Method</th>
<th>#Modules</th>
<th>Precision (±δ)</th>
<th>Recall (±δ)</th>
<th>F-measure (±δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DivAfull</td>
<td>28</td>
<td><strong>0.56</strong> (±0.44)</td>
<td><strong>0.46</strong> (±0.40)</td>
<td><strong>0.49</strong> (±0.40)</td>
</tr>
<tr>
<td>MaWish</td>
<td>24</td>
<td>0.50 (±0.48)</td>
<td>0.38 (±0.40)</td>
<td>0.41 (±0.42)</td>
</tr>
<tr>
<td>DivAfull</td>
<td>18</td>
<td>0.87 (±0.12)</td>
<td><strong>0.71</strong> (±0.25)</td>
<td><strong>0.76</strong> (±0.21)</td>
</tr>
<tr>
<td>MaWish</td>
<td>13</td>
<td><strong>0.93</strong> (±0.10)</td>
<td>0.70 (±0.27)</td>
<td><strong>0.76</strong> (±0.22)</td>
</tr>
</tbody>
</table>

Table 5.2: The average of precisions, the average of recalls, and the average of F-measures of worm protein modules. The upper part reports results for all complexes, the bottom part for all functional predictors.

have a higher average precision but a better recall is obtained by DivAfull modules. However, in total, they are of the same quality as shown by the average of F-measures (the bottom part of Table 5.2).

<table>
<thead>
<tr>
<th>Species</th>
<th>#Functions</th>
<th>FNR</th>
<th>#Predictors</th>
<th>CVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast</td>
<td>144</td>
<td>0.23</td>
<td>39</td>
<td>0.26</td>
</tr>
<tr>
<td>worm</td>
<td>90</td>
<td>0.06</td>
<td>18</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 5.3: The total number of biological functions predicted by DivAfull functional predictors and their functional ratio and the total number of DivAfull functional predictors and their coverage ratio computed with respect to MaWish results.

Furthermore, it is interesting to investigate whether DivAfull modules also provide new predictions. By computing functional and coverage ratio over all functions predicted by DivAfull functional predictions with respect to biological functions of MaWish predictions, Table 5.3 shows that there is a particular fraction of new discoveries for both species.

To sum up, we may conclude that DivAfull discovered a higher number of conserved complexes of the comparable or higher quality than MaWish. DivAfull also achieved new predictions.

5.5.3 Divide and MNAligner

MNAligner applies MCODE to each sub-graph produced by Divide before using the alignment procedure. Despite of the high number of generated
sub-graphs generated by Divide, many of them have an empty set of complexes detected by MCODE. Indeed, the final number of conserved complexes is low, 12 complexes for Saccharomyces cerevisiae and 10 modules for Caenorhabditis elegans. However, almost the same number of complexes is discovered when MCODE is directly applied on orthologous sub-networks of the species being compared (see Tables 5.4 and 5.5, respectively). These results seem to indicate that the low number of discovered complexes is due to characteristics of MCODE’s clustering approach.

Tables 5.4 and 5.5 show the average of precisions, the average of recalls, and the average of F-measures of the detected complexes for Saccharomyces cerevisiae and Caenorhabditis elegans, respectively, after measuring their GO enrichment.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Modules</th>
<th>Precision (±δ)</th>
<th>Recall (±δ)</th>
<th>F-measure (±δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divide+MNAligner</td>
<td>12</td>
<td>0.74 (±0.25)</td>
<td>0.53 (±0.28)</td>
<td>0.60 (±0.27)</td>
</tr>
<tr>
<td>MNAligner</td>
<td>13</td>
<td>0.69 (±0.24)</td>
<td>0.48 (±0.27)</td>
<td>0.53 (±0.28)</td>
</tr>
<tr>
<td>Divide+MNAligner</td>
<td>11</td>
<td>0.81 (±0.09)</td>
<td>0.58 (±0.24)</td>
<td>0.65 (±0.21)</td>
</tr>
<tr>
<td>MNAligner</td>
<td>12</td>
<td>0.75 (±0.12)</td>
<td>0.52 (±0.24)</td>
<td>0.58 (±0.24)</td>
</tr>
</tbody>
</table>

Table 5.4: MNAligner: The average of precisions, the average of recalls, and the average of F-measures of yeast protein modules. The upper part reports results for all complexes, the bottom part for all functional predictors.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Modules</th>
<th>Precision (±δ)</th>
<th>Recall (±δ)</th>
<th>F-measure (±δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divide+MNAligner</td>
<td>10</td>
<td>0.72 (±0.39)</td>
<td>0.55 (±0.4)</td>
<td>0.59 (±0.37)</td>
</tr>
<tr>
<td>MNAligner</td>
<td>11</td>
<td>0.38 (±0.45)</td>
<td>0.34 (±0.43)</td>
<td>0.35 (±0.43)</td>
</tr>
<tr>
<td>Divide+MNAligner</td>
<td>8</td>
<td>0.90 (±0.11)</td>
<td>0.68 (±0.31)</td>
<td>0.74 (±0.23)</td>
</tr>
<tr>
<td>MNAligner</td>
<td>5</td>
<td>0.83 (±0.15)</td>
<td>0.75 (±0.29)</td>
<td>0.77 (±0.24)</td>
</tr>
</tbody>
</table>

Table 5.5: MNAligner: The average of precisions, the average of recalls, and the average of F-measures of worm protein modules. The upper part reports results for all complexes, the bottom part for all functional predictors.

From Table 5.4 it can be seen that the complexes of Saccharomyces cerevisiae discovered when Divide was applied, and their subset of functional predictions, outperformed the complexes and predictions of the straightforward
ward application of MNAligner (with MCODE).

For Caenorhabditis elegans, if we consider all modules, again better results are achieved when Divide is incorporated prior the clustering and alignment steps (the upper part of Table 5.5). When we focused on functional predictions, the application of Divide lead to results of higher precision but lower recall, which also affected the F-measure (the bottom part of Table 5.5). However, from 10 conserved complexes discovered when Divide is applied, 8 are potential functional predictions, but, in the case when orthologous sub-networks are not divided, more than the half of the results do not satisfy criteria for functional prediction.

<table>
<thead>
<tr>
<th>Species</th>
<th>#Functions</th>
<th>FNR</th>
<th>#Predictors</th>
<th>CVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast</td>
<td>109</td>
<td>0.28</td>
<td>11</td>
<td>0.36</td>
</tr>
<tr>
<td>worm</td>
<td>48</td>
<td>0.46</td>
<td>8</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 5.6: The total number of predicted biological functions and their functional ratio and the total number of functional predictors and their coverage ratio as result when Divide is combined with (MCODE and) MNAligner computed with respect to the results of straightforward application of (MCODE and) MNAligner.

In the end, we computed functional and coverage ratio over all functions and their functional predictions detected with the method which includes Divide with respect to the results of the application of MCODE and MNAligner. Table 5.6 indicates that in both species a quarter or even more of the results are new discoveries.

In summary, the application of Divide resulted in new and in the majority of the cases better results despite the fact that the same clustering technique was applied on Divide sub-graphs as on the original whole orthologous sub-networks before the division. This shows that Divide can positively bias the search for improving detection of conserved complexes by means of modular global network alignment.
5.6 Conclusions

The *Divide* method can be used to perform network alignment of protein interaction networks by acting on pairs of the identified sub-graphs (Jancura et al., 2008a,b). In particular, we tested experimentally the ability of *Divide* to be used for performing modular network alignment. Specifically, we performed two comparative experimental analysis.

In the first experiment we used the *DivAfull* algorithm, which uses *Divide* prior to the alignment phase, as done by Jancura et al. (2008a). Comparison between results of *MaWish* and *DivAfull* indicated that *DivAfull* is able to discover new alignments which significantly increase the number of discovered complexes. Moreover, complexes discovered by *DivAfull* showed comparable or improved GO enrichment, as measured by precision, recall, and F-measure, and provided new prediction of protein functions. This application shows that using *Divide* one can enhance the search strategy by replacing greedy with exact search in the alignment graph, resulting in the discovery of new conserved complexes.

In the second experiment an instance of global network alignment approach, called *MNAligner*, was considered. This method employs a pre-processing step before computing the alignment of two PPI networks. The results showed that the application of *Divide* enhanced the quality of the results. This indicates the regions around articulation hubs constructed by *Divide* provide a beneficial search bias for detecting functional complexes and enhancing the performance of *MNAligner*.

In summary these results showed that *Divide* can be successfully applied to discover conserved protein complexes and to ‘refine’ state-of-the-art algorithms for network alignment.
Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


128


Bibliography


Bibliography


Samenvatting

De vooruitgang van laboratorium- en computationele technieken voor het ontdekken van fysieke bindingen en functionele afhankelijkheden tussen eiwitten heeft het mogelijk gemaakt om grote hoeveelheden gegevens over eiwitinteracties in verschillende organismen te verzamelen. Deze verzameling van eiwit-eiwit interacties (PPIs) kan worden gepresenteerd in een grote graaf, een PPI netwerk, waar de eiwitten de knopen van de graaf zijn en de eiwit-eiwit interacties de kanten tussen de knopen. Een dergelijk netwerk model heeft veel eigenschappen en interne structuren die gerelateerd zijn aan biologische functies en organisatie niveaus. Analyse van de topologische kenmerken van een PPI netwerk is dus een krachtige manier om de organisatie en functies van cellulaire componenten in levende wezens te begrijpen.

In het algemeen worden eiwitten in een organisme geproduceerd door de overdracht van sequentiële informatie gecodeerd in de genen. Dit betekent dat de functionaliteit en de interacties van eiwitten worden beïnvloed door de evolutie van het genoom. Men kan dus sporen van deze evolutie verwachten in de topologie van PPI netwerken. Kennis van evolutionair behoud en divergente van de interne structuur van een PPI netwerk kan nieuwe inzichten opleveren in evolutionaire processen en levende systemen. Daarom is evolutionaire analyse in PPI netwerken een van de centrale onderzoekslijnen geworden binnen de systeembiologie.

Dit proefschrift beschrijft bijdragen op het gebied van evolutionaire analyse in PPI netwerken. Na een korte inleiding wordt in hoofdstuk 2 een uitgebreide literatuurstudie gegeven over dit onderwerp. We onderscheiden
twee hoofdrichtingen van onderzoek. De ene groep onderzoeken beschouwt een PPI netwerk van een enkele soort en hoe de topologie en functionele eiwitmodules beperkt worden door de evolutie. De tweede groep van onderzoeken richt zich op het evolutionair behoud en divergentie in meerdere PPI netwerken en de inferentie van biologische informatie over verschillende soorten. Dit hoofdstuk beschrijft ook succesvolle toepassingen van de evolutionaire benaderingen voor netwerkanalyse en hun toekomstperspectieven.

In hoofdstuk 3 beschouwen we vervolgens een enkel PPI netwerk en hoe de aanwezigheid van evolutionair geconserveerde eiwitten het vinden van complexen in dit netwerk beinvloed. De voorgestelde methode kwantificeert de evolutionaire bias van geconserveerde eiwitten met de biologische functies van evolutionair-gedreven eiwitcomplexen en toont aan dat de evolutionair-gedreven eiwitcomplexen duidelijk te onderscheiden zijn van de niet evolutionair-gedreven complexen.

Hoofdstuk 4 beschouwt opnieuw een PPI netwerk van een enkele soort en beschrijft een algoritme dat het netwerk onttrekt in subnetwerken, zodat deze subnetwerken evolutionair geconserveerde complexen bevatten. Het verschil tussen evolutionair geconserveerde complexen en de evolutie-gedreven complexen in hoofdstuk 3 is dat evolutionair geconserveerde complexen een tegenhanger moeten hebben in het netwerk van een andere soort, terwijl dit voor evolutie-gedreven complexen niet het geval hoeft te zijn. De uitgebreide experimentele analyse toont aan dat we de evolutionair geconserveerde complexen in een PPI netwerk lokaal kunnen bestuderen.

Ten slotte beschouwt het laatste hoofdstuk, hoofdstuk 5, twee PPI netwerken van verschillende soorten en vergelijkt deze door gebruik te maken van het ontbindingsalgoritme uit hoofdstuk 4. Om precies te zijn wordt elk PPI netwerk voorbewerkt met dat algoritme, waarna een vergelijkingsmethode wordt toegepast. Experimenten laten zien dat met het ontbindingsalgoritme de resultaten van state of the art netwerkvergelijkingmethoden wordt verbeterd. Verder demonstreren zij een proof of concept van een modulaire manier om PPI netwerken te vergelijken.
Summary

The advancement of laboratory and computational techniques for uncovering physical bonds or functional interdependencies between proteins has enabled the accumulation of a large amount of protein interactome data for various species. The set of protein-protein interactions (PPIs) can be represented by a large-sized graph, called PPI network, where proteins are the nodes of the graph and protein-protein interactions are the edges between the respective nodes of the graph. Such network model may exhibit many properties and inner structures that can relate to various biological functions and organizational levels. Thus, analysis of PPI networks and its topological features is a powerful approach to understanding the biological organization and the function of cellular components in living organism.

In general, proteins in an organism are produced by the transfer of sequential information encoded by genes. Consequently, proteins’ functionality and the way they interact are influenced by the genome evolution. Hence, one may expect the presence of evolutionary imprints in the topology of PPI networks. Knowing evolutionary conservation and divergence of PPI network’s inner structures can bring new insights into understanding evolutionary processes and living systems as such. Therefore, evolutionary analysis in PPI networks has become one of the central research areas in systems biology.

This thesis presents particular contributions to the field of evolutionary analysis in PPI networks. After a brief introduction, Chapter 2 provides an extensive literature overview on the topic. Two main directions of the research are distinguished, where one group of studies considers a
PPI network of a single species and how its topology and functional protein modules are constrained by the evolution, while the second group of studies focuses on the evolutionary conservation and divergence in multiple PPI networks and the inference of biological information across various species. The chapter also discusses successful applications of evolutionary approaches for network analysis and their future perspectives.

Next, in Chapter 3 we study a single PPI network and how the presence of evolutionary conserved proteins affects the detection of protein complexes in the network. The proposed methodology quantifies the evolutionary bias of conserved proteins with the biological functions of evolutionary-driven protein complexes and shows that the evolutionary-driven protein complexes are clearly differentiated from the non-evolutionary-driven ones.

Chapter 4 again considers a single PPI network of one species and describes an algorithm which decomposes the network into sub-networks such that these sub-networks contain evolutionary conserved complexes. The difference between evolutionary conserved complexes and evolutionary-driven complexes in Chapter 3 is that evolutionary conserved complexes must have their evolutionary counterparts in the network of another species, while evolutionary-driven complexes need not satisfy this constrain. The extensive experimental analyses demonstrate that one can study the evolutionary conserved complexes in PPI networks locally.

Finally, the last chapter, Chapter 5, considers two PPI networks of different species and performs their comparison by exploiting the divisive algorithm proposed in Chapter 4. Specifically, each PPI network is preprocessed with that algorithm prior applying any comparative method. Experimental results show the ability of the divisive algorithm to improve the performance of state-of-the-art network comparison methods and demonstrate a proof-of-concept for performing a PPI network comparison in a modular fashion.
Curriculum vitae

Pavol Jancura was born on 1. October 1983 in Svidník in former Czechoslovakia, currently Slovakia, and he grew up in a small village Malá Pol’ana. After finishing the secondary school Gymnasium DH in Svidník in 2002, he studied 4 out of 5 years a master track in computer science at Faculty of Mathematics and Physics at Charles University in Prague. At the end of the 4th year he successfully applied for the one year short-track master programme and the university scholarship at Faculty of Science at Vrije University in Amsterdam, where he completed the 5th year of his master studies and obtained the master’s degree in computer science in July 2007. Then he started his doctoral studies at Centre for Integrative Bioinformatics at Vrije University. As his main supervisor later took a job position at Radboud University in Nijmegen and Pavol wanted to continue his research under the same guidance, he decided to follow his supervisor and at the end of January 2008 moved to Nijmegen to join Intelligent Systems group at Radboud University. The main results of his research work he conducted until 2012 are summarized in this thesis.
SIKS Dissertatiereeks

2009

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models

2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques

2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT

2009-04 Josephine Nabukonya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-06 Muhammad Subianto (UU)
Understanding Classification

2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems

2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications

2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
Operating Guidelines for Services

2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems
2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies
(making ontologies work in e-science with ONTO-SOA)

2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess

2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data

2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System

2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated
Electronic Markets

2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controlling Influences on Decision Making

2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification

2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence

2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment

2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations

2009-25 Alex van Ballegooij (CWI)
RAM: Array Database Management through Relational Mapping

2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27 Christian Glahn (OU)
Contextual Support of Social Engagement and Reflection on the Web

2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage

2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?

2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks

2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks

2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings

2009-42 Toine Bogers (UvT)
Recommender Systems for Social Bookmarking

2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients
2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations

2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful

2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

====

2010
====

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter

2010-02 Ingo Wassink (UT)
Work flows in Life Science

2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia Documents

2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI

2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance

2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions

2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines
2010-27 Marten Voulon (UL)
    Automatisch contracteren

2010-28 Arne Koopman (UU)
    Characteristic Relational Patterns

2010-29 Stratos Idreos (CWI)
    Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT)
    Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31 Victor de Boer (UVA)
    Ontology Enrichment from Heterogeneous Sources on the Web

2010-32 Marcel Hiel (UvT)
    An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33 Robin Aly (UT)
    Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34 Teduh Dirgahayu (UT)
    Interaction Design in Service Compositions

2010-35 Dolf Trieschnigg (UT)
    Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36 Jose Janssen (OU)
    Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37 Niels Lohmann (TUE)
    Correctness of services and their composition

2010-38 Dirk Fahland (TUE)
    From Scenarios to components

2010-39 Ghazanfar Farooq Siddiqui (VU)
    Integrative modeling of emotions in virtual agents

2010-40 Mark van Assem (VU)
    Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
    Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services

2010-46 Vincent Piipers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions

2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives

2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources

2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

====

2011

====

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier (UU)
Organizing Agent Organizations. Syntax and Operational Semantics of an
Organization-Oriented Programming Language

2011-03 Jan Martijn van der Werf (TUE)
   Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU)
   Insights in Reinforcement Learning: Formal analysis and empirical evaluation of
temporal-difference learning algorithms

2011-05 Base van der Raadt (VU)
   Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging
   Discipline.

2011-06 Yiwen Wang (TUE)
   Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT)
   Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU)
   BDI-based Generation of Robust Task-Oriented Dialogues

2011-09 Tim de Jong (OU)
   Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT)
   Cloud Content Contention

2011-11 Dhaval Vyas (UT)
   Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE)
   Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT)
   Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR)
   Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA)
   The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
   Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA)
   Exploring Topic Structure: Coherence, Diversity and Relatedness
2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU)
The Mind's Eye on Personal Profiles

2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein (UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification

2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science
2011-33 Tom van der Weide (UU)  
Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU)  
Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU)  
Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU)  
Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN)  
Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM)  
Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU)  
Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU)  
Architectural Knowledge Management in Global Software Development

2011-41 Luan Ibraimi (UT)  
Cryptographically Enforced Distributed Data Access Control

2011-42 Michal Sindlar (UU)  
Explaining Behavior through Mental State Attribution

2011-43 Henk van der Schuur (UU)  
Process Improvement through Software Operation Knowledge

2011-44 Boris Reuderink (UT)  
Robust Brain-Computer Interfaces

2011-45 Herman Stehouwer (UvT)  
Statistical Language Models for Alternative Sequence Selection

2011-46 Beibei Hu (TUD)  
Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

2011-47 Azizi Bin Ab Aziz (VU)  
Exploring Computational Models for Intelligent Support of Persons with Depression

2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

2012-01 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda

2012-02 Muhammad Umair (VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

2012-03 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories

2012-04 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications

2012-05 Marijn Plomp (UU)
Maturing Interorganisational Information Systems

2012-06 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks

2012-07 Rianne van Lambalgen (VU)
When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

2012-08 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories

2012-09 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms

2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment

2012-11 J.C.B. Rantham Prabhakara (TUE)
Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems
2012-13 Suleman Shahid (UvT)
Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

2012-14 Evgeny Knutov (TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems

2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes

2012-16 Fienke Both (VU)
Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance

2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices

2012-19 Helen Schonenberg (TUE)
What’s Next? Operational Support for Business Process Execution

2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval

2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval

2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

2012-26 Emile de Maat (UVA)
Making Sense of Legal Text

2012-27 Hayrettin Gurkok (UT)
Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
2012-28 Nancy Pascall (UvT)
   Engendering Technology Empowering Women

2012-29 Almer Tigelaar (UT)
   Peer-to-Peer Information Retrieval

2012-30 Alina Pommeranz (TUD)
   Designing Human-Centered Systems for Reflective Decision Making

2012-31 Emily Bagarukayo (RUN)
   A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

2012-32 Wietske Visser (TUD)
   Qualitative multi-criteria preference representation and reasoning

2012-33 Rory Sie (OU)
   Coalitions in Cooperation Networks (COCOON)