NOTE TO USERS

This reproduction is the best copy available.

UMI®

PIPE: A PROTEIN–PROTEIN INTERACTION PREDICTION ENGINE BASED ON THE RE-OCCURRING SHORT POLYPEPTIDE SEQUENCES BETWEEN KNOWN INTERACTING PROTEIN PAIRS

by

Sylvain Pitre

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

School of Computer Science

at

CARLETON UNIVERSITY

Ottawa, Ontario
January, 2010

© Copyright by Sylvain Pitre, 2010

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-63858-3
Our file Notre référence
ISBN: 978-0-494-63858-3

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Identification of protein interaction networks has received considerable attention in the post–genomic era. The currently available biochemical approaches used to detect protein–protein interactions (PPIs) are all time and labor intensive. Consequently there is a growing need for the development of computational tools that are capable of effectively identifying such interactions.

In this thesis we explain the development and implementation of a novel Protein–Protein Interaction Prediction Engine termed PIPE. This tool is capable of predicting protein–protein interactions for any target protein pair of the yeast *Saccharomyces cerevisiae* from their primary structure and without the need for any additional information or predictions about the proteins. PIPE showed a sensitivity of 61% for detecting any yeast protein interaction with 89% specificity and an overall accuracy of 75%. PIPE was used to identify novel interactions and a novel yeast complex confirmed by tandem affinity purification (TAP tag).

We also report an improved version, PIPE2, which exhibits a specificity of approximately 99.95% and executes 16,000 times faster than the original method. Importantly, we report an all–to–all sequence–based computational screen of PPIs in yeast in which we identify 29,589 high confidence interactions out of approximately 2 x 10^7 possible pairs. Furthermore, a novel putative protein complex was discovered, comprised largely of membrane proteins not amenable to TAP tagging.

The third iteration of the PIPE algorithm (PIPE3) has been adapted and tested to predict interactions in several organisms including *H. sapiens* and involving several viruses. A computational genome-wide scan of *S. pombe* revealed over 9,000 PPIs, triple what was available using traditional methods. The novel PPIs in *S. pombe* were used to identify two new complexes which have been studied in more detail. PIPE3 has also been shown to be able to make cross-organism predictions: predicting PPIs in one organism using PPI knowledge of one or more other organisms. These tests suggest that PIPE3 can be a good predictor for newly sequenced organisms or for those which have very few known PPIs.

Acknowledgements

First, I would like to thank my supervisor Dr. Frank Dehne for his guidance, help an support in the development and improvement of this thesis. Special thanks to Dr. Ashkan Golshani for his biological expertise as well as his experimental validations of our computational predictions. Thanks to Dr. James Cheetham for his ideas and input during the creation of the original PIPE. I would also like to thank Dr. James R. Green and Dr. Michel Dumontier for their insights, discussions and invaluable feedback on the development of PIPE2. Thanks to Chris North for his algorithmic and programming improvements to our first version of PIPE. Thanks to Xuemei Luo and Ryan Taylor for their programming expertise during the early stages of this project. Last but not least thanks to all the people involved in research and biological experiments throughout this project: M. Alamgir, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Hooshyar, M. Jessulat and N. Krogan.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	Amino Acid
AD	Activating Domain
AUC	Area Under the Curve
DBD	DNA Binding Domain
DNA	DeoxyriboNucleic Acid
DSB	Double–Stranded DNA Break
GO	Gene Ontology
HIV	Human Immunodeficiency Virus
LC–MS	Liquid Chromatography–Mass Spectrometry
MS	Mass Spectrometry
NHEJ	Non–Homologous End–Joining
ORF	Open Reading Frame
PPI	Protein–Protein Interaction
RNA	RiboNucleic Acid
ROC	Receiver Operating Characteristic
SDS-PAGE	Sodium Dodecyl Sulfate–PolyAcrylamide Gel Electrophoresis
SVM	Support Vector Machine
TAP	Tandem Affinity Purification
TEV	Tobacco Etch Virus
UAS	Upstream Activating Sequence
Y2H	Yeast Two–Hybrid

Chapter 1

Introduction

1.1 Introduction

Proteins carry out the majority of the biological processes in cells. Most often, proteins accomplish this task in association with protein partners, forming stable or transient protein complexes. It is therefore generally accepted that protein–protein interactions are responsible for the cell's behavior and its responses to various stimuli [10, 88, 109]. Further, the completion of higher eukaryotic genome projects have led to the understanding that the biological complexity underlying higher organisms is not accomplished by increasing the number of genes [1, 63, 132]. It is now thought that this complexity stems from an elevated pattern of protein–protein interactions in higher organisms [22, 103]. As a consequence, charting protein–protein interaction maps remains a major goal in biological research. A large part of post–genomic research has focused on the analysis of protein–protein interactions. Measurement, prediction and analysis of interactions between proteins have been extensively used to identify proteins that are functionally related. As a consequence, analysis of protein interaction networks has become a powerful tool to assign putative functions to previously ill–characterized proteins [10, 109]. In this context, the yeast *Saccharomyces cerevisiae* has emerged as the model organism for studying functional proteomics. In the past, protein interaction analysis has been used to assign putative functions to different yeast proteins [59, 135].

Protein–protein interactions can be most readily identified by protein affinity chromatography or pull-down experiments, yeast two–hybrid screens (Y2H), or purifying protein complexes that have been tagged *in vivo*. These methods are all labour and time consuming and have a high cost associated with them. Each of them has inherent advantages and disadvantages. The yeast two-hybrid system has the advantage of identifying the direct interaction between protein pairs [51, 80]. However, the data

gathered from this method has a high rate of false positives (as much as 40%) and in the absence of other lines of evidence, this data alone may not be considered as biologically significant [28, 29, 114]. Affinity purification methods such as the *in vivo* double–tagging of protein complexes followed by purification steps using affinity chromatography, also known as tandem affinity purification (TAP tag), has the advantage of identifying complexes that really exist *in vivo* (as long as the tagged protein is not overproduced) [99, 101]. However, all affinity purification methods suffer from limitations [28, 29, 35]. First, the addition of a tag, large or small, to the protein may change its properties, causing changes in complex stability or composition. Second, all purification methods suffer from the copurification of "contaminating" proteins. It is often difficult to conclude whether these "contaminants" represent true endogenous partners or artificial associations induced by cell disruption. Third, during affinity purifications proteins are isolated as complexes and therefore the direct interactions between protein pairs are not readily distinguished from the indirect (via intermediates) ones. The high cost, as well as the technical limitations associated with such biochemical approaches has resulted in a growing need for the development of computational tools that are capable of identifying protein–protein interactions.

1.2 Motivation

The high cost, as well as the technical limitations associated with traditional biochemical approaches, has yielded a growing need for the development of computational tools that are capable of identifying protein–protein interactions. As a result, there have been a number of such tools developed over the past few years. Some of these tools are based on previously identified domains [44, 56, 117], some use similarities and sequence conservation between interacting proteins [32, 50], others use the structural information of proteins [4, 5, 85]. The primary structure of the proteins has also been used to detect protein–protein interactions. Using a vector based learning machine it has been shown that the primary sequence of amino acids alone may successfully be used to detect protein–protein interactions [16, 77]. A disadvantage of the protein–protein interaction detection tools is that they often have limited abilities to detect novel interactions and to differentiate them from false positives. A high rate

of false negatives is another disadvantage associated with some of these tools.

1.3 Statement of the Problem

Due to the drawbacks of traditional methods there is the need for accurate computational prediction of protein–protein interactions using currently available information. Some methods use protein structure information or are based on a set of previously detected domains but that information is far from complete for any organism. However, for most organisms of interest, the complete sequenced genome and therefore the primary structure (sequence) for every protein are available. For yeast in particular, there exists multiple databases of tested and predicted interactions. Consequently a method which uses the sequence information only as input and which uses the available interactions already discovered by traditional methods to make accurate predictions would be of great interest.

1.4 Contributions

This thesis presents a new algorithm for protein–protein interaction prediction based on the re-occurring short polypeptide sequences between known interacting protein pairs called PIPE: **P**rotein–protein **I**nteraction **P**rediction **E**ngine [94]. This algorithm makes it possible to detect protein–protein interactions in *S. cerevisiae* using only sequence information (primary structure) as opposed to methods that rely on physical structure or known motifs or domains. Also this computational method is not restricted to the limitations of *in vivo* testing (i.e. some proteins cannot be tested successfully *in vivo*). It has been used to predict novel interactions and a novel process for which the internal structure has been confirmed by tandem affinity purification (TAP tag). A PIPE portal is available at http://cgmlab.carleton.ca/PIPE/.

An improved version of the algorithm called PIPE2 is also presented [95]. PIPE2 brings drastic speed improvement over PIPE along with higher specificity at the expense of lower sensitivity. The increased speedup of the algorithm is due to optimized window comparisons, pre-computation and query approach. A modified version of the

median filter is used on the results to improve the specificity which decreases the expected number of false positives significantly. However the filter has the negative, but expected, side–effect of reducing the sensitivity. With the increased speed it was possible to do a genome-wide scan of the *S. cerevisiae* genome by running all possible protein pairs ($\approx$ 20 million pairs). The PIPE2 interaction result list is larger than recent large–scale experiments using traditional methods. A PIPE2 portal is available at http://pipe.cgmlab.org/ along with executable binaries and our complete dataset.

Finally, the latest version of the algorithm named PIPE3 demonstrates how the PIPE algorithm can be applied to other organisms such as *C. elegans, S. pombe, E.coli* but most importantly *H. sapiens* (human). This expanded version is shown to successfully predict PPIs in other organisms but is also shown to be able to predict PPIs in organisms using other organisms as sources of information. The important implication of this feature is the ability to predict PPIs in new or unstudied organisms even given the lack of known interactions. PIPE3 has been used to run a genome–wide scan of the *S. pombe* organism as well as new PPI predictions in *H. sapiens, Hepatitis C, Influenza A, HIV-1* and *HIV-2*. It is demonstrated to be a good predictor for newly sequenced organisms or for those which have very few known PPIs due to the fact PIPE3 can make predictions using a collection of PPIs from multiple organisms. A PIPE3 portal is available at http://cgmlab.carleton.ca/PIPE3/.

1.5 Overview of Results

The original PIPE has been able to predict protein–protein interactions in *S. cerevisiae* with an estimated sensitivity of 61% and specificity of 89%. It was successful in predicting a novel interaction between proteins YGL227W–YMR135C which was subsequently verified by TAP–tagging. PIPE has also been demonstrated to detect the sites of interactions for a known interaction. From a small pool of 10 interacting pairs, PIPE was able to detect the correct site of interaction in 4 of them. This method was used to detect a novel protein complex termed vid30c (YIL017C, YMR135C, YDL176W, YIL097W, YBR105C and YDR255C) but more importantly it was able to show the internal structure of this complex. The predicted interactions within the complex were also confirmed experimentally by TAP–tagging. The PIPE

algorithm has recently been independently reviewed and compared favorably against other similar sequence–based methods [89].

PIPE2 is an improvement in both speed and specificity of the PIPE algorithm. Combining the optimized window comparisons, pre-computation and query approach a speedup of more than 16,000 times compared to the original algorithm is achieved. The increased specificity obtained by the modified median filter is estimated at 99.95% (compared to 89% for the original PIPE). PIPE2 was used to scan the entire *S. cerevisiae* genome ($\approx 6,400$ proteins) which includes ≈ 20 million possible protein pairs. This experiment produced a list of 29,589 predicted interactions of which nearly half (14,438 or 48.8%) are not found in any other database (novel interactions). This predicted list is larger than other recent large–scale experiments using traditional methods (Gavin *et al.* [36], Krogan *et al.* [61]) as well as other comparable computational methods (Betel *et al.* [15], Wang *et al.* [131]). A sample of these predictions were validated by hand and 66% were supported by at least one line of evidence (function or 3^{rd} party interaction).

A novel complex was discovered within the PIPE2 novel interactions list which consists of 8 proteins: YGL051W, YAR027W, YAR028W, YCR007C, YAR033W, YOR307C, YLR065C and YKL174. According to the study of the members of this complex it is likely to be a real complex in the *S. ceverisiae* organism. The predicted PIPE2 interactions were compared to other large scale traditional and computational experiments to illustrate the overlap between various methods. Also the PIPE2 interactions along with other traditional large–scale results were grouped according to GO Slim locations. Our method is shown to predict significantly more membrane interactions than the previous traditional methods.

The latest version of PIPE, called PIPE3, has been extended to predict PPIs in many other organisms such as *C. elegans, S. pombe, E.coli* and *H. sapiens.* The sensitivity/specificity of PIPE3 is evaluated and measured in each organism in order gauge PIPE's accuracy. Hundreds of novel human PPIs are predicted and analyzed, predictions have been experimentally verified and confirmed, PPIs between human proteins and viral pathogen proteins are identified, the potential interaction sites of proteins are determined, and an all-to-all PPI interaction map for *S. pombe* (9,000

PPIs) was predicted.

Using predicted novel PPIs from PIPE3 two novel protein complexes were identified in *S. pombe*. One of them is a five member complex with SPBC1289.13c, a putative galactosyltransferase as a core protein that interacts with four other proteins, some of which interact with each other. The second complex consists of 8 members, with 5 proteins forming the core, and 3 additional proteins that interact with the core proteins but not with each other. Five of these proteins (SPBC1703.10, SPAC4C5.02c, SPAC6F6.15, SPAC9E9.07C and SPAC18G6.03) have been linked to protein transport and vesicular trafficking.

1.6 Organization of Thesis

First we will present protein–protein interaction basics and background in *Chapter 2 – Background*. We will then review traditional methods for protein–protein interactions as well as survey previous computational methods that are available in *Chapter 3 – Literature Review*. Next we will introduce the protein–protein interaction engine (or PIPE) used to detect novel interactions in yeast in *Chapter 4 – PIPE: Protein–protein Interaction Prediction Engine*. The following chapter, *Chapter 5 – PIPE2: Improving PIPE*, will offer an improved version of the PIPE algorithm called PIPE2 which increases the accuracy and decreases the runtime significantly which allows us to run the entire yeast genome to produce a complete interaction map for this organism. In *Chapter 6 – PIPE3: Predictions in Other Organisms* will discuss the most recent version of PIPE adapted to run on several other organisms. The results of these experiments will be explained and some of them will be experimentally verified to confirm our findings. We will also show the web–portals of all three PIPE versions as well as their settings in *Chapter 7 – PIPE Web–Portal*. Finally in *Chapter 8 – Conclusion* we will offer a summary of the results, our closing remarks as well as future work.

1.7 Summary

We have defined the protein–protein interaction problem and discussed why we need to solve it, as well as an overview of the contributions and results. Let us now present some detailed background information on protein–protein interactions.

Chapter 2

Background

2.1 Sensitivity and Specificity

Throughout this thesis we will refer to measures of quality for experimental and computational protein–protein interactions. These measures, sensitivity and specificity, are explained here so that the reader can understand the discussions that follow.

Sensitivity is calculated as:

$$\frac{TP}{TP+FN}$$

Specificity as:

$$\frac{TN}{TN+FP}$$

and accuracy as:

$$\frac{TP+TN}{TP+FN+FP+TN}$$

expressed in percentage (%) where TP is the number of true positive, FN the number of false negatives, TN the number of true negatives, and FP the number of false positives. In simplified terms, sensitivity measures the capability of a method to detect true positives as positives and specificity measures the capability of a method to detect true negatives as negatives. Ideally, one would want both of these numbers to be as high as possible.

2.2 Introduction

Proteins are key biomolecules that often realize their functions by interacting with one another. Protein–protein interactions (PPIs) mediate various aspects in the structural and functional organization of a cell including multi–faceted responses to internal

and external stimuli. Protein interaction networks have also been shown to possess topological and dynamic properties that may be essential for certain biological events [44, 53]. Thus, elucidating the complete network of PPIs is expected to garner a greater understanding of the biology of the cell.

The sequencing of the budding yeast *Saccharomyces cerevisiae* over a decade ago [39] has led to its emergence as the model organism of choice for large-scale functional genomics experiments including expression profiling [38] and identification of PPI networks (interactomes) using yeast-two hybrid screens [51, 127]. The early PPI investigations had little overlap and neither of them detected more than 13% of the published interactions detected through conventional analyses [47]. Recent investigations using the Tandem Affinity Purification (TAP) followed by mass spectrometry have generated higher confidence yeast protein interactomes [61, 119]. This approach does not require protein over-expression and is expected to yield the interactions that really exist *in vivo* [35].

However, the lack of significant overlap between the data collected in different genome-wide TAP tag experiments and those collected using yeast two–hybrid [29] together with the experimental limitations associated with each of these techniques suggests that the interaction networks are far from being saturated [46, 49]. Consequently, there is a growing need for the development of new and improved experimental and computational approaches to better uncover the yeast interactome.

Recently, we [94] as well as others [65] reported that PPIs could be successfully detected from short polypeptide sequences within proteins. Our approach that we termed Protein–protein Interaction Prediction Engine, PIPE, is based on re–occuring short polypeptide sequences observed in a database of known interacting protein pairs. With a sensitivity of 61%, specificity of 89%, we demonstrated how PIPE can predict novel and existing PPIs for any target pair of yeast proteins from their primary structure alone, regardless of their physical or biological properties. The use of protein sequence alone was an improvement over traditional approaches which are often restricted by the properties of certain proteins. For example, it is often difficult to investigate membrane proteins using experimental approaches.

Previous genome–wide analyses of PPIs have predominantly relied on Y2H and

TAP tag methodologies. These techniques are both time and labor intensive and they both have high rates of false positives and false negatives results associated with them ($\approx$ 40% false positive rate for Y2H [114] and 15–50% false positive rate for TAP tag [29]). For example, like all affinity purification methods, TAP tag suffers from co–purification of contaminants. Additionally, it is often difficult to differentiate between direct and indirect (i.e. via a third partner) interactions using TAP tag. Another common limitation for these techniques is that they cannot be applied to all proteins without discrimination. In TAP tag, the double tag fusion to the target protein may interfere with the formation of some complexes or cause a mutant phenotype [35, 133]. In Y2H, not all proteins can be safely over–expressed and not all proteins can find their way into the nucleus which is required for the successful detection via Y2H [58]. The goal of PIPE is to complement previous genome–wide experimental analyses of PPIs, leading to a more complete PPI map.

2.3 Protein–Protein Interactions

An overwhelming number of biological processes are mediated through the action of proteins. In many cases, these proteins carry out their functions by interacting with each other in either stable or transient protein complexes. The nature and increasing complexity of these interactions is thought to be responsible for the overall biological complexity in higher organisms. Therefore, it is believed that humans, for example, are more sophisticated than the nematode *C. elegans*, not only because we possess marginally greater number of genes, but largely because human proteins form more intricate networks [3, 22]. Recent advances in the field of genomics and proteomics have led to the discovery and characterization of some of these networks [63, 128]. An organism may have numerous interactomes representing different tissue types, biological states, etc. The complete elucidation of all interaction networks found in an organism will have significant implications for science [97]. For example, the cellular roles and molecular functions for previously ill–characterized proteins may be inferred from the networks of interactions that they participate in. Moreover, the conservation of protein interactomes across organisms will also provide insight into their evolutionary relationships.

Practically, knowledge of interaction networks will provide insight into their dependencies and lead to enhanced approaches for drug discovery. For these reasons, the elucidation of protein–protein interactions (PPIs), especially within the context of an interaction network, is an important goal in biological research [34, 105]. Until recently, PPIs were determined by carrying out experiments that were specifically designed to identify a small number of specifically targeted interactions. However, the development of novel genomic techniques allows for high–throughput experiments, which can now be carried out to exhaustively probe all possible interactions within an entire genome. *S. cerevisiae*, also known as Baker's yeast, has emerged as the model organism of choice for functional proteomics due to the elucidation of its genomic sequence in 1996 [39]. Since then, whole PPI maps have been determined using various methods including yeast two–hybrid [51, 127], affinity purification/mass spectrometric identification methods such as TAP–tagging [48, 72], and protein chips [122, 142]. Indirect large–scale approaches such as synthetic lethal analysis [122] and correlated mRNA expression profile [37] have also been used to investigate PPIs.

However, these methods are not without shortcomings. Not only are they labor– and time–intensive, they also have a high cost associated with them. Another important disadvantage is the poor accuracy of the high–throughput data generated. Significant discrepancies between results of small–scale high–confidence experiments and high–throughput studies have been reported [4, 34]. Interstudy discrepancy is even higher when comparing data generated from different large–scale studies [4, 34]. In addition, the PPI data obtained from biological experiments often include many false positives, which may connect proteins that are not necessarily related and should be confirmed by other methods. Consequently, there is a growing need for the development of computational tools that are capable of effectively identifying PPIs as well as interpreting and validating the experimentally derived data.

2.4 Summary

Now that we have introduced some background information on protein–protein interactions, we will present a literature review which will describe in details the traditional and computation methods of detecting protein–protein interactions.

Chapter 3

Literature review

The book chapter published by Pitre et al. [93] is the basis for this chapter.

3.1 Introduction

A wide range of computational methods have been developed to build, study, and exploit protein interactomes (reviewed in [4, 34, 62, 69, 93, 113, 115, 139]). First, computational methods have been developed to construct interaction databases within which experimentally determined data is collected and annotated. Automated data mining techniques can then be applied to extract relevant information about potential interactions from the vast amount of PPI information in these databases. As mentioned earlier, a number of experimental techniques have been used to determine large–scale protein interaction maps. Although the significant inconsistencies between interaction maps of the same organism obtained using different techniques can be somewhat justified [34], computational methods have been successfully applied to assess, validate, and carefully scrutinize these experimentally determined protein interactomes. Based on the assumption that physically interacting proteins have a high probability of also being functionally related, a number of computational tools have been developed to exploit protein interaction networks in order to predict functional features of the proteins. Lastly, computational methods can also be used to predict novel PPIs by learning from known interactions [4, 34, 62, 69, 93, 113, 115, 139].

It is the objective of this chapter to provide an overview of these computational methods, with the main focus being on computational tools for the prediction of novel interactions. We also highlight the specific limitations for each of the tools discussed, as well as the systematic shortcomings common to most computational tools. For comparison, the advantages and limitations of traditional "wet lab" experimental approaches are also summarized. Finally, due to the large number and frequency of

new PPI methods released, it is impossible to include all tools relevant to the study of PPIs and the author apologize in advance to all those researchers whose work has not been cited here. While not every method in existence is listed in this chapter, any approach should place most other methods in one (or more) of the listed categories.

3.2 Traditional PPI Prediction Methods

Traditional PPI prediction methods are those done *in vivo* in a wet lab. While there are several methods to detect PPI *in vivo* we will direct our attention to two of the most popular: Yeast–Two Hybrid (Y2H) and Tandem Affinity Purification (TAP).

3.2.1 Yeast–Two Hybrid (Y2H)

The yeast two–hybrid (Y2H) method was one of the first methods to be applied to the detection of PPIs. Two protein domains are required in the Y2H assay that have specific functions: (*i*) a DNA binding domain (DBD) that helps bind to DNA, and (*ii*) an activation domain (AD) responsible for activating transcription of DNA. Both domains are required for the transcription of a reporter gene [51]. The Y2H assay relies on the fusion of DBD to a protein of interest (X) at its N–terminus and the fusion of AD to another protein of interest (Y) at the C–terminus, which forms DBD–X (bait) and AD–Y (prey). If the bait and prey hybrids interact with each other, the transcription of the reporter gene will be induced and, in this way, the interaction can be detected [33]. Y2H analysis allows the direct recognition of PPI between protein pairs. However, a large number of false positive interactions may arise, while a number of true interactions will be missed (i.e., false negatives). A false positive interaction can occur by activation of RNA polymerase by a bait protein, by the binding of the prey AD–Y protein with upstream activating sequences (UAS), by non–specific binding of bait and prey proteins with some endogenous proteins, or by the binding of "sticky" prey proteins with bait proteins [120]. On the other hand, many true interactions may not be detected using Y2H assay, leading to false negative results. In a Y2H assay, the interacting proteins must be localized to the nucleus; since membrane proteins are typically less likely to be present in the nucleus they are unavailable to activate reporter genes, and hence are excluded. Proteins

that require post–translational modifications to carry out functions are also unlikely to behave or interact normally in a Y2H experiment. Furthermore, if the proteins are not in their natural physiological environment, they may not be folded properly to interact [110]. During the last decade, Y2H has been improved by designing new yeast strains containing multiple reporter genes and new expression vectors to facilitate the transformation of yeast cells with hybrid proteins [101].

3.2.2 Tandem Affinity Purification (TAP)

Tandem affinity purification (TAP) tagging was developed to study PPIs under the native conditions of the cell [101]. Gavin *et al.* first attempted the TAP–tagging method in a high–throughput manner to analyze the yeast interactome [35]. This method is based on the double tagging of the protein of interest on its chromosomal locus, followed by a two–step purification procedure using Staphylococcus protein A and calmodulin beads separated by a tobacco etch virus (TEV) protease cleavage site. First, a target protein open reading frame (ORF) is fused with the DNA sequences encoding the TAP tag and is expressed in yeast where it can form native complexes with other proteins. The tagged protein, along with its associated proteins/complexes, is then extracted from the cell lysate. The fused protein and the associated complexes are then purified via a two–step affinity purification procedure. Proteins that remain associated with the target protein can then be analyzed and identified through SDS–PAGE [102] followed by mass spectrometry analysis [101], thereby identifying the PPI partner proteins of the original protein of interest. An important advantage of TAP–tagging is its ability to identify a wide variety of protein complexes and to test the activity of monomeric or multimeric protein complexes that exist *in vivo*. Compared to Y2H, TAP–tagging obtains interaction information from a more natural environment since the physiological conditions are more realistic than those created by Y2H, including factors like post–translational modifications and pH requirements. However, the TAP tag may interfere with the formation of some protein complexes (as shown by [35]) by low expression of fusion proteins [104], which can affect the ability of a protein to interact with other proteins or may cause a mutant phenotype [133]. These problems may be minimized by using other complementary

techniques that can increase the reproducibility of any large–scale approaches.

3.2.3 Interaction Databases

The large quantity of experimental PPI data being generated on a continual basis necessitates the construction of computer–readable biological databases in order to organize and effectively disseminate this data. A number of such databases exist and are growing at exponential rates. The Biomolecular Interaction Network Database (BIND), for example, is built on an extensible specification system that permits detailed description of the manner in which the PPI data was derived experimentally, often including links directly to the supporting evidence from the literature [9]. The database of interacting proteins (DIP) is another database of experimentally determined protein–protein binary interactions [136]. DIP serves as an access point to a number of other related databases such as LiveDIP, which provides information on the functional aspects of protein complexes as well as links out to other databases such as the database of ligand–receptor partners (DLRP). The General Repository of Interaction Datasets (BioGRID) is a database that contains genetic and physical protein–protein interactions among proteins from 13 species [119]. Interactions are regularly added through exhaustive curation of the primary literature. Interaction data is extracted from other databases including BIND and MIPS (Munich Information center for Protein Sequences) [79], as well as directly from large–scale experiments [20]. The molecular interaction database (MINT) is another database of experimentally derived PPI data extracted from the literature, with the added feature of providing the weight of evidence for each interaction [20]. The Human Protein Reference Database (HPRD) offers a centralized database of curated *H. sapiens* protein–protein interactions from several sources [55]. There are several other databases dedicated to single organisms (such as EciD *E. coli* [6]) or other large collections of experiments, but only the ones discussed in this thesis are listed here.

3.3 Computational PPI Prediction Methods

Computational methods provide a complementary approach to detecting PPIs. Indeed, the wide availability of experimental data has spurned the development of numerous computational methods over the past few years. In general, all computational approaches to PPI prediction attempt to leverage knowledge of experimentally determined previously known interactions in order to predict new PPIs. These methods enable one to discover novel putative interactions and often provide information for designing new experiments for specific protein sets. These approaches can be classified into five general categories: methods based on genomic information, evolutionary relationships, three dimensional protein structure, protein domains, and primary protein structure. Specific approaches that fall within these categories are listed in Table 3.1 and are discussed below. Figure 3.1(a)–(e) presents the idea behind the five categories of methods. Other methods such as the PRINCESS [67] combine several of these methods and known information about the proteins to assess the quality of protein–protein interactions. For example the PRINCESS method uses orthologous interactions, interacting domains, GO annotation, gene coexpression, genome context and network topology to assess the reliability of human protein–protein interactions from high–throughput experiments.

3.3.1 Genomic Methods

Genomic methods for interaction prediction take advantage of the availability of information obtained by complete genome sequencing. Completely sequenced genomes provide knowledge of which genes are present and how they are organized (gene order). The conservation of gene order across species yields information about the evolution of the genome, and hints at which genes may be functionally correlated. Most computational methods that use genomic information do not rely solely on the sequence similarity between homologous genes (or their products) [30, 73], but rather assess functional links between pairs or clusters of co–located genes. Evidence for the evolutionary conservation of gene order can be obtained by systematic comparison of completely sequenced genomes. Dandekhar *et al.* [25] compared nine bacterial and archaeal genomes and applied a method based on co–localization to determine

Method	Description
Whole genome	Conservation of gene across genomes [25]. Comparison of protein pairs in one genome to its fused single protein product homolog in another genome [31, 74].
Evolutionary relashionship	Correlated evolution of functionally related proteins [91]. Tree kernel-based computational system to assess similarities between phylogenetic profiles [90, 129].
3D protein structure	Assess fit of two interacting partners on a predetermined complex of known 3D structure; InterPreTS [5, 71]. Multimeric threading algorithm MULTIPROSPECTOR to recognize partners in protein interactions [70]. CAPRI is a community-wide experiment focusing on the performance of protein–protein docking procedures [134]. PRISM: protein interactions by structural matching [85].
Domain based	Combination of similarity between sequence patches involved in interactions and between domains of interacting partners [32]. Maximum likelihood estimation method to determine probability of interactions between evolutionarily conserved protein domains in the Pfam protein domain database [26]. Prediction of interaction probability of proteins; ranking system for probability of interactions between protein pairs [44, 45]. Database of potentially interacting domain (PID) pairs using a DIP database and InterPro; PID matrix score as a reliability index for accurate analysis of interaction networks [56]. GAIA by Zhang *et al.* [141]
Primary protein structure	Protein interactions mediated through specific short polypeptide sequences [117, 139]. Automatic recognition of correlated patterns of sequences and substructure by support vector machine; also uses associated physiochemical parameters [16]. Combination of sequence information, experimental data analysis and subsequence paring to generate a signature product that is implemented with support vector machine (SVM) [77]. Kernel methods for predicting protein–protein interactions [12]. Prediction using SVM with a kernel function and sequence information [112]. Structure-templated predictions of novel protein interactions from sequence information [15]. InSite: protein–protein interaction binding sides on a proteome-wide scale [131]. Codon usage to predict protein–protein interaction using sequence information [82]. PIPE: protein–protein interaction prediction using primary protein structure data from MIPS and DIP databases [94]

Table 3.1: Computational methods for the prediction of protein–protein interaction.

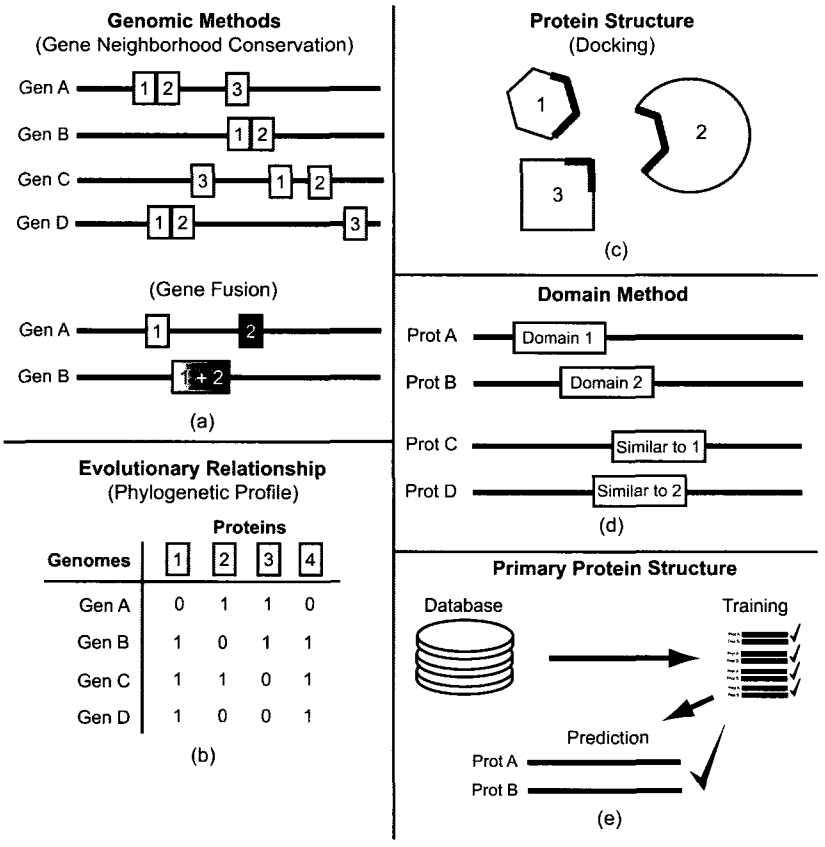

Figure 3.1: The five categories of computation PPI methods: (a) genomic methods, (b) evolutionary relationship, (c) protein structure, (d) domain method and (e) primary protein structure.

conserved gene pairs even within relatively low conservation of gene–order. They found that proteins encoded by conserved gene pairs also tend to interact physically. Physical interactions between encoded proteins have been demonstrated for at least 75% of the conserved gene pairs. A further 20% of the conserved pairs were predicted to encode proteins that interact physically [25]. While promising, the approach fails to identify interactions between products of distantly located genes. Moreover, false predictions are generated because the proximity constraint is not sufficient to determine physical interaction. Finally, this approach may not be applicable to eukaryotes, because the co–regulation of genes is not imposed at the genome structure level [74].

The co–localization of genes encoding interacting or functionally related gene

products can be taken a step further. Pairs of interacting or functionally related proteins sometimes have homologs in another genome in which they are fused into a single protein [75]. For example, the Gyr A and Gyr B subunits of *Escherichia coli* DNA gyrase are fused as a single protein in yeast topoisomerase II [74]. Thus, the sequence similarities between Gyr A and Gyr B and different segments of the topoisomerase II might be used to predict that Gyr A and Gyr B may interact in *E. coli* [74]. It is important to note however that there are not many examples of gene fusions. Marcotte *et al.* developed a computational method to search for such fusion events within multiple genomes. In their study, they uncovered 45,502 such putative PPIs in yeast. Some proteins that were found to be linked to several other proteins also appeared to interact functionally in pathways. Many of these putative interactions were also confirmed experimentally, as documented in the DIP database. Similarly, Enright *et al.* identified 215 genes involved in 64 unique fusion events across *E. coli*, *Haemophilus infuenzae* and *Methanococcus jannaschii* [31]. This gene–fusion analysis approach has since been incorporated into a computational algorithm for the prediction of PPIs and protein function [75].

3.3.2 Evolutionary Relationship Methods

Evolutionary relationships between two proteins can also be used to infer a physical and functional relationship. The phylogenetic profile of a protein describes the presence of homologs across a series of organisms. Proteins that exhibit similar profiles may be functionally linked. For instance, proteins that make up multimeric structural complexes or that participate in a given biochemical pathway typically exhibit similar phylogenetic profiles. Pellegrini *et al.* applied phylogenetic profiling to predict the function of previously uncharacterized proteins [91]. The comparison of profiles is further enhanced by including evolutionary information. Vert showed that the accuracy of function prediction using a support vector machine (SVM) is improved with the use of evolutionarily enhanced phylogenetic profiles [129]. A comparative genome phylogenetic analysis approach has also led to prediction of hundreds of pairs of interactions in *E. coli*, and thousands in yeast [90].

3.3.3 Protein Structure Methods

As the number of experimentally solved protein structures continues to increase, three–dimensional (3D) structure information has become increasingly applied to the prediction of physical binding [4, 70]. By considering homologous proteins, it has been shown that close homologs (> 30% sequence identity) physically interact in the same or similar way [4]. Aloy and Russell describe such a 3D–based method to model putative interactions [4]. The method assesses the fit of two potential interacting partners on a complex of known 3D structure and infers molecular details of how the interaction is likely to occur. In general, it has been shown that residues located at the interface tend to be structurally conserved [71]. Residues that make atomic contacts in a crystallographic complex are analyzed. An interaction is conserved as long as the contacting residues is also conserved. Homologs of both interacting proteins are then examined to see whether these interactions are preserved. All possible pairs between two protein families can then be modeled and the most likely interactions determined. The method also offers the means of assessing the compatibility of a proposed PPI within such a complex, as well as for ranking interacting pairs in studies that involve protein families that show different interaction specificities. The method can be used to model a complex based on the known structure of a similar template complex, and to correctly predict interactions within several systems [4]. Aloy *et al.* successfully demonstrated how 3D structures can be used to query entire interaction networks so as to validate and infer the molecular details of interactions that have been predicted using other methods.

CAPRI (Critical Assessment of PRedicted Interactions) is a community–wide experiment that aims to fairly evaluate the state of the art in protein–protein docking procedures by making predictions on a set of interacting proteins for which the solution has not yet been published [134]. Models are compared to high quality crystallographic interaction data by independent CAPRI assessors. During the course of these experiments, it was found that models exhibiting a high degree of native intermolecular contacts were generally good indicators of true PPIs. PRISM (Protein Interactions by Structural Matching) searches a dataset of protein structures for potential interaction partners by comparing protein structure pairs with a dataset of

interfaces [85]. This interface dataset is a structurally and evolutionary representative subset of biological and crystal interactions present in the PDB. The algorithm calculates the similarity between interfaces by first obtaining structural surface alignments. This measures structural similarity of a target structure to a binding site. If the surfaces of two target proteins contain similar regions to complementary partner chains, it may be inferred that those target proteins interact through similar regions. The PRISM web server allows users to explore protein interfaces as well as predictions of PPIs. One can search a variety of stored interfaces categorized by functional clusters or structural similarity. For example, users can search for proteins involved in cell metabolism, while restricting the results to interfaces of certain sizes. PRISM's interactive visualization tool shows the 3D model along with the desired features. One can also submit protein structures (in PDB format) for interaction prediction. Note that this method is only applicable to proteins with known structure.

InterPreTS (interaction prediction through tertiary structure) is a web–based version of the above method [5]. Homologs of a test pair of protein sequence are identified from the database of interacting domains (DBID) of known 3D complex structures. The sequences are then scored for how well they preserve sites of contacts at the interaction interface [5]. InterPreTS allows one to visualize the molecular details of any predicted interaction. Combining domain structural similarities and conserved sequence patches among interacting proteins has also led to improved methods for interaction prediction [32]. Lu *et al.* report a multimeric threading approach to identifying interaction partners and to assign quaternary structures of proteins found in the yeast DIP database [70]. This multimeric threading algorithm, MULTIPROSPECTOR, is able to recognize partners involved in protein interactions and correctly predict a significant number of interacting yeast proteins pairs that have already been identified in the DIP database. The method correctly recognized and assigned 36 of 40 homodimers, 15 of 15 heterodimers, and 65 of 69 monomers that were scanned against a protein library of 2,478 structures obtained from the Protein Data Bank (PDB) [14]. The reported prediction accuracy of current methods often varies substantially, and recent efforts have been made to address this issue.

3.3.4 Domain Based Methods

There are a number of computational techniques that are based solely on the conservation of protein domains. For example, a method developed by Deng *et al.*, employs maximum likelihood estimation to infer interacting domains that are consistent with the observed PPIs [26]. Using evolutionary conserved domains defined in the Pfam (protein families) protein domain database [116], the probabilities of interactions between every pair of domains are estimated. These inferred domain–domain interactions are subsequently used to predict interactions between proteins. Han *et al.* provide a similar computational tool that not only predicts the PPIs, but also provides the interaction probability of input proteins and ranks the possibilities of interaction between multiple protein pairs [44, 45].

Another prediction algorithm called PreSPI (prediction system for protein interaction), based on conserved domain–domain interactions, was also described by Han *et al.* [44]. Here a domain combination–based PPI probabilistic framework is used to interpret PPIs as the result of interactions of multiple domain pairs or of groups. This tool is able to predict the interaction probability of proteins and also provides an interaction possibility ranking method for multiple protein pairs that can be used to determine which protein pairs are most likely to interact with each other in multiple protein pairs. A high sensitivity of 77% and specificity 95% were obtained for the test groups containing common domains when tested using an interacting set of protein pairs found in the yeast DIP database. Correlations were observed between the interacting probability and the accuracy of the prediction, making the output probability a useful indicator of prediction confidence. This method was also somewhat successful when tested on an artificially made random pairing of proteins used as a negative test set of non–interacting protein pairs. This method is particularly advantageous because it also allows for mass prediction of genome–wide interactions, which in turn makes it possible to construct entire protein interaction networks.

Kim *et al.* developed a database for potentially interacting domain pairs (PID) refined from the DIP database of interacting proteins by making use of InterPro, an integrated database of protein families, domains, and functional sites [56]. A statistical scoring system, "PID matrix score" was developed as a reliability index for

accurate functional analysis of interaction networks and a measure of the interaction probability between domains. This method combines various kinds of information such as sequences, interacting regions, and domains of both interacting partners [56]. In order to evaluate the predictive power of the PID matrix, cross–validation was performed with subsets of DIP data (positive datasets) and randomly generated protein pairs from TrEMBL/SwissProt database (negative datasets). The prediction system resulted in approximately 50% sensitivity and more than 98% specificity [56]. The result also showed that mapping of the genome–wide interaction network can be achieved by using the PID matrix.

Finally, Zhang *et al.* presents a novel approach called Gram-bAsed Interaction Analysis (GAIA, [141]) to predict PPIs in *S. cerevisiae*. It is based on n–grams, or short strings of n amino acids, from annotated domain–domain interaction data. This method achieves a true positive rate of 82% and a false positive rate of 21% on a set of gold–standard interactions. However like other domain–based methods, GAIA will not be able to predict interactions mediated by amino acids that are not part of a currently known interacting domain.

3.3.5 Primary Protein Structure Methods

Primary protein structure approaches are predicated on the hypothesis that PPIs may be mediated through a specific number of short polypeptide sequences. These sequences do not span whole domains but are found repeatedly within the proteins of the cell. SVM–based learning methods have shown that the primary sequence of an amino acid chain can effectively identify PPIs [16, 77].

Spriznak *et al.* presents an approach that integrates the predictions obtained from different computational approaches together with experimental data, so as to provide functional assignments [117]. It was reported that characteristic pairs of sequence–signatures can be learned from a database of experimentally determined interacting proteins, where one protein contains the first sequence signature and its interacting partner contains the other sequence–signature. The sequence–signatures that appear together in interacting protein pairs are termed correlated sequence–signatures. This analysis is applied to a database of experimentally identified interacting protein pairs

in yeast, from which distinct over–represented sequence–signature pairs were identified. Although not every protein with the one signature is expected to interact with every protein with the other signature, this approach can be used to direct and narrow down experimental interaction screens [117].

Another approach is based on the ability of an SVM learning system to automatically recognize correlated patterns of sequence and substructure in the interacting pairs of proteins found in the DIP database. These patterns typically comprise a small number of functional residues in each protein. This computational tool, developed by Bock and Gough, is based on primary structure information as well as associated physico–chemical properties such as charge, hydrophobicity, and surface tension. Reported prediction accuracy was 80%, but the test set size was very small (five previously characterized interactions) [16]. Martin *et al.* describe an algorithm for PPI prediction [77] that follows the approach of Bock and Gough by combining sequence information and experimental data analysis, while extending the concept of sequence–signatures from Sprinzak *et al.* by using subsequence pairing. Information from experimental data, sequence analysis, and local descriptions of protein pairs, which are more representative of the actual biology of PPI, are combined to generate a novel and even more general descriptor called a signature product. The signature product is then implemented within a SVM classifier as a kernel function [77]. This method was applied to publicly available yeast datasets among others. The yeast and *H. pylori* datasets used to verify the predictive ability of the method yielded accuracies of 70–80% using tenfold cross–validation. The human and mouse datasets were also used to demonstrate that the method is capable of cross–species prediction. This method is advantageous over that of Bock and Gough because it uses only experimental and sequence information, and does not require physico–chemical information. In addition, this approach, unlike that of Sprinzak *et al.*, does not require prior knowledge of domains.

Ben–Hur and Noble [12] also make use of SVMs to predict PPIs, but introduce a novel pair–wise kernel that measures the similarity between two pairs of proteins. SVMs and kernel methods have the ability to integrate different types of information through the kernel function. Here, kernels make use of a combination of data including

protein sequence, homologous interactions, and GO annotations. Ben–Hur and Noble explore a number of different kernel functions using yeast PPI data from the BIND database. At a false positive rate of approximately 1%, the sensitivity was 80%. Future directions may include data incorporation from gene expression studies and transcription factor binding data that have been useful in predicting PPIs. A recent paper by Shen *et al.* [112] presents another method based on a SVM with a kernel function using only sequence information to predict PPI in humans. The authors report an average prediction accuracy of 83.90%. Recently, Deng *et al.* also proposed a similar method based on SVM

Betel *et al.* [15] have published a method based on structure templates to predict novel protein interactions from sequence information. They rely on existing physical interaction data in order to build sequence profiles that determine the binding specificity of interaction domains. This method is called domain-motif interactions from structural topology (D-MIST). They publish a total of 18,459 interactions between 2,313 proteins in *S. cerevisiae.* They note that 609 interactions within that set have reported experimental evidence ($\approx 3\%$) but no sensitivity or specificity are given for the results.

Wang *et al.* [131] have developed a tool for identifying protein interaction binding sites called InSite. Their method uses a library of conserved motifs, a dataset of protein interactions and indirect evidence of protein–protein and motif–motif interactions (i.e. expression correlation, Gene Ontology, annotation, domain fusion) to predict binding sites on a proteome-wide scale. This technique uses the different input information to find possible explanations for the predicted binding sites. They publish an interaction dataset of $\approx 80,000$ interactions involving $\approx 2,700$ proteins in *S. cerevisiae.* The Guo *et al.* [43] approach also uses motifs but attempts to discover very short motif pairs (3–8 amino acids) responsible for interactions between protein pairs and uses them to predict PPIs.

A method by Najafabadi *et al.* [82] uses codon usage to predict protein–protein interactions using only the primary protein structure. This method called 'PIC' (probabilistic–interactome using codon usage) combines the information provided by the frequencies of all codons by way of a naïve Bayesian network in order to predict

protein–protein interactions in *S. cerevisiae*. The authors claim an increase of 75% in specificity at a precision of 50% compared to predictions which do not consider codon usage.

The method discussed in this thesis called PIPE (protein–protein interaction prediction engine) is able to predict with high confidence PPIs for any target pair of yeast proteins given only knowledge of their primary structure data [94]. Like other PPI prediction methods, PIPE relies on previously acquired experimentally derived PPI data and extrapolates this information to predict novel PPIs. This engine compiled the dataset of 15,118 PPI pairs of *S. cerevisiae* from the DIP [136] and MIPS [79] databases. PIPE predicts the probability of interaction between two proteins by measuring how often pairs of subsequences in two query proteins A and B are observed to co–occur in pairs of protein sequences known to interact. PIPE showed an overall accuracy of 75%, a success rate that is on par with other commonly used biochemical techniques. PIPE analysis also has other applications in that it can be used to study the internal architecture of yeast protein complexes [94]. It has been independently assessed [89] against other methods previously discussed by Shen *et al.* [112], Martin *et al.* [77] and Guo *et al.* [43].

A new version of the PIPE method, called PIPE2, has been created in order to run genome-wide predictions in *S. cerevisiae* [95]. This thesis will also present an extension of the PIPE algorithm (PIPE3) to predict PPIs in multiple organisms. This multi-organism PPI prediction is something few other computational methods have attempted to do, instead only targeting one organism.

A method called Linear Motif Discovery (LMD) contains some parallel features to PIPE [83]. In that report the primary sequences of proteins in the database of interacting protein pairs were analyzed to identify novel protein interaction motifs. In this manner the authors identified dozens of novel interacting motif candidates. A significant difference between PIPE and this approach is that PIPE is optimized to predict the likelihood of an interaction between a given pair of proteins, whereas LMD is optimized to identify protein–protein binding motifs. The existence of a protein–protein binding motif in a pair of proteins does not indicate how likely this is going to result in an actual protein–protein interaction.

Zaki *et al.* [139] have developed a method called PPI–PS (Protein–protein Interaction based on Pairwise Similarity) which is very similar in most respect to PIPE. This method however concatenates all the sequences into one long vector and typically uses much larger windows than PIPE (500-20,000AA as opposed to 20AA in PIPE). It also uses a Smith-Waterman score which is the score of the best local alignment with gaps between a subsequence and all protein sequences. Using SVM, PPI–PS is claimed to attain 77.89% accuracy (80.7% sensitivity at 74.4% specificity) when tested on 4,917 interacting and 4,000 non-interacting pairs.

3.3.6 Strengths, Weaknesses, and Challenges of Computational PPI Predictions

Researchers have embraced the use of computational methods in the elucidation of PPIs. Computational PPI prediction methods are an invaluable source of information that complement labor–intensive experimental approaches such as Y2H and TAP–tagging. However, the high–throughput nature of bioinformatics tools should require that computational predictions be deemed reliable only after proper scrutiny. Appropriate measures to evaluate the significance of the interactions should be developed to minimize the number of results that give false positives and negatives. While it is often difficult to differentiate between novel interactions and false positives, additional contextual clues including function, expression, and localization should be brought into consideration. As computational methods are based directly or indirectly on experimentally obtained data, the inaccuracies in the original data will likely be propagated into the predictions.

Several other factors contribute to the challenges that face computational PPI predictions. False positives are prevalent in most computational methods, but we can easily find an explanation. The model organism used for testing in many methods, yeast, contains roughly 6,300 proteins [42], which yields approximately $\approx$ 20 million possible pairs. Even with a false positive rate as low as 1%, we would anticipate $\approx 200,000$ falsely predicted interactions. It has been estimated that, in actuality, there are anywhere between 10,000 and 30,000 interactions in yeast [11, 42, 47, 64, 118, 124, 130]. Recent large–scale studies contain datasets of a size closer to the bottom end of that

range (7,123 in Krogan *et al.* [61]). We can therefore see that the positive interactions are vastly outnumbered by the number of negative interactions. Even if we assume there are 30,000 possible interactions there is still more than a 600:1 ratio of negative to positive interactions ($\approx 0.158\%$). Therefore it is extremely difficult to recognize the true positive predictions among the overwhelming background of false positive predictions. The lack of reliable a gold standard makes the assessment of prediction accuracy by the various tools somewhat arbitrary. The establishment of a gold standard is essential to measure progress in the field and will also serve as training material for the next generation of prediction methodologies. Strong gold standard datasets need to be constructed from multiple lines of evidence, including structure where possible, and made freely available. Recent developments in computational interaction prediction have opened the door to predicting entire interactomes for a variety of organisms. For the most sophisticated approaches, this objective is very computationally expensive and time–consuming. However, algorithmic optimizations and continued improvements in hardware performance will help overcome these challenges.

3.4 Summary

In spite of the number of challenges that are faced in the use of computational methods, one can only expect that they will have even wider applications in the genome–wide analysis of interactomes. The most obvious result of this will be the enlargement of protein databases. It is also expected that the efficiency of these methods will improve. At present, there is an emergence of a more integrated strategy in which genomic, proteomic, and other forms of data are incorporated into the process of generating protein interaction maps. It appears that these strategies will also be able to take other cellular processes such as post–translational protein modification and protein degradation into consideration. It is impossible to deny the invaluable insight into the organization of living organisms that has been provided by even the simplest of protein interaction models. As these models become more sophisticated, computational methods will become of higher importance.

Chapter 4

PIPE: Protein–protein Interaction Prediction Engine

4.1 Introduction

We have previously discussed the traditional methods such as Y2H and TAP-tag along with their strengths and weaknesses. They are generally time consuming and costly methods with low accuracy. These drawbacks motivated development of computational protein–protein interactions predictions. However each computational method also has strengths and weaknesses. Some methods require the 3D structure of each protein in order to make a prediction. Unfortunately, since we don't have the structure for a lot of proteins in yeast those proteins, interactions cannot be predicted using those methods. Methods based on known motifs suffer from a similar problem: what if your motif database is incomplete or incorrect? There is therefore the need for a simple method that uses as little information about the protein as possible.

In this chapter, we report on the development and implementation of a computational tool termed Protein–Protein Interaction Prediction Engine (PIPE). This engine uses the primary structure of proteins together with the available protein interaction data to predict the potential interaction between any target pairs of *S. cerevisiae* proteins. A paper published by Pitre et al. [94] is partly the basis for this chapter.

4.2 The Problem

Here we ask the question: can novel protein–protein interactions be successfully predicted from amino acid sequences (the primary structures) alone and without any further information/prediction about the proteins? Our hypothesis is that some of the interactions between proteins are mediated by a finite number of short polypeptide sequences. These sequences may be typically shorter than the classical domains and are used repeatedly in different proteins and in different contexts within the cell.

Once the interaction database is large enough to sample these sequences, it should be possible to accurately predict such protein–protein interactions. We therefore need an algorithm that will only use an interaction database as its source of information and that can predict an interaction between two proteins given only their primary structure (or sequence).

4.3 Algorithm Description

Our protein–protein interaction prediction algorithm relies on previously determined interactions. At the time we initiated this study, our dataset was composed of 15,118 pairs of protein–protein interactions using a total of 6,304 yeast protein sequences and was compiled from the *S. cerevisiae* protein interactions reported in the DIP [106] and MIPS [79] databases. These interactions were determined using several methods, each having a limited accuracy. Since our algorithm is based on uncertain data, we expect a certain degree of error associated with our predictions. The principle of our method is as follows: assume we have two query proteins A and B, along with the knowledge that certain proteins C and D are interacting. If a region (subsequence) a_1 in A resembles a region in C, and a sequence b_1 in B resembles a region in D, there is a possibility that A and B are also interacting via an interaction between the corresponding a_1 and b_1 subsequences, which co–occur in both protein pairs $A - B$ and $C - D$. As the number of interacting protein pairs in the database which contain the corresponding sequences a_1 and b_1 increases so does the likelihood that a_1 and b_1 are the true mediators of an interaction between A and B.

To match two windows or subsequences, PIPE uses a substitution matrix. Substitution (or scoring) matrices describes the rate at which one character in a sequence changes to other character states over time. This 20×20 matrix will contain values for every pair of amino acid according to how likely one could change into the other due to evolution or mutation. After some trial and error we settled on the PAM120 substitution matrix. The PAM matrix stands for Point Accepted Mutation, in which PAM1 would be based on probabilities of 1 accepted mutation per 100 amino acids (1%). PAM120 therefore represents 120 accepted mutations per 100 amino acid. One should use higher–numbered PAM matrices to compare more evolutionary distant

sequences.

The algorithm can be divided into the following steps (see also Figure 4.1):

Step 1: Input the dataset of known protein interactions (referred to as the interaction list):

(a) Every protein in the interaction list is represented as a node in graph G.

(b) Every interacting pair of sequences in the interaction list results in an edge l in G between the two respective nodes.

(c) Input two query sequences: A of length m and B of length n.

Step 2: Sequence A is fragmented into overlapping segments of w amino acids each. In other words, we use a sliding window of length w and move it forward by one amino acid in each step. For each fragment a_i, $i = 1$ to $(m - w + 1)$, we do the following:

(a) Search for fragment a_i in every sequence in the database. We also use a sliding window of length w in every sequence in the database. For every fragment in each sequence we use a substitution matrix (PAM120) to match the corresponding amino acids with a_i. We define a score which is the sum of PAM120 scores for the w pairs of amino acids matched. That score will be used to identify whether two fragments are similar or not.

(b) For every sequence containing a fragment that matches a_i (score equal or greater than a threshold S_{pam}), we add to a list R all neighbors of that sequence in G (by following its adjacent edges in G).

Step 3: Once all fragments a_i have been searched in the database and all neighbors of successful matches have been added to the list R, we search all fragments b_j of sequence B in R. As in Step 2, we use again a sliding window of size w to create fragments b_j, $j = 1$ to $(n - w + 1)$, of B and then search each b_j in R. Every match of a b_j in R will result in a score increment of one in a result matrix where each row i represents a fragment a_i in A and each column j represents each fragment b_j in B.

Step 4: The result matrix is presented as a 3D surface where the rows and columns represent the fragments a_i and b_j, respectively, and the elevation represents the score S, i.e. the number of matches observed for the corresponding fragments a_i and b_j.

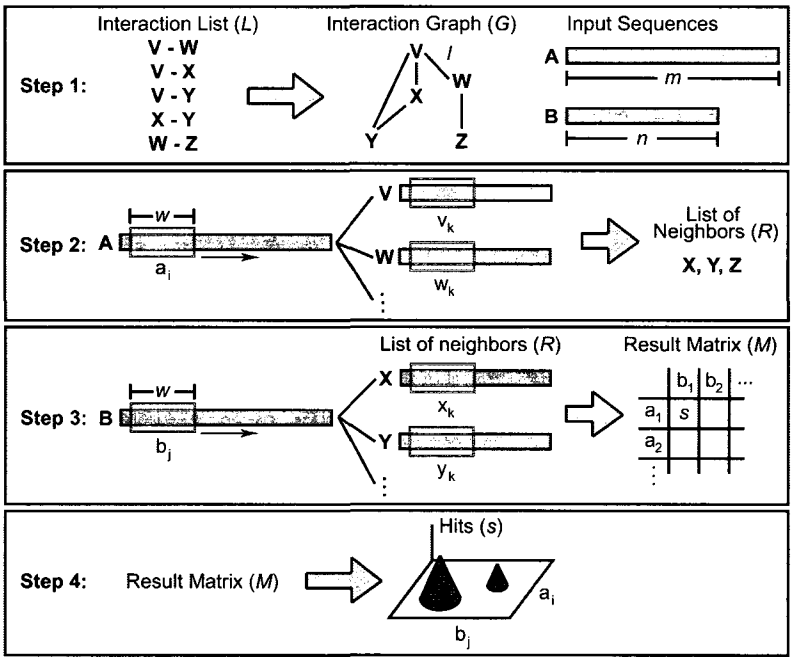

Figure 4.1: Illustration of PIPE algorithm. Step 1: input the sequences (A, B) and build an interaction graph from the interaction list. Step 2: scan the interaction graph for proteins matching A and build a list of neighbors. Step 3: scan the list of neighbors for proteins matching B and for every hit increment the corresponding cell in the result matrix. Step 4: graph the result matrix into a 3D surface where a hill of height greater than 10 indicates an interaction.

4.4 Algorithm Pseudocode and Analysis

The PIPE algorithm was described informally in the previous section so we present here the pseudocode description. $|P|$ represents the length in amino acids (AA) of protein P or the number of proteins in list or graph —P— depending on the context. The input consists two query proteins A and B as well as the interaction graph G. The output will consist of the $|A| \times |B|$ matrix H.

procedure *PIPE*(G, H, A, B)

```
1: Input graph G
2: Input query proteins A of length m and B of length n
3: for every a_i in A of length w for i = 0 to (|A| − w) do
4:   for every protein V in graph G do
5:     for every v_j of length w in V for j = 0 to (|V| − w) do
6:       if compare(a_i, v_j) ≥ S_pam then
7:         create a list R containing every neighbor of V in G
8:         for for every protein X in R do
9:           for every x_t of length w in X for t = 0 to (|X| − w) do
10:            for every b_u in B of length w for u = 0 to (|B| − w) do
11:              if compare(b_u, x_t) ≥ S_pam then
12:                H[i][u] = H[i][u]+1
13:                stop comparing again X and go to next neighbor in R
14:              end if
15:            end for
16:          end for
17:        end for
18:      end if
19:    end for
20:  end for
21: end for
22: return H
```

From the pseudocode we can analyze the complexity of the PIPE algorithm to determine an upper–bound on the worst–case runtime. It is clear that the runtime of the algorithm will depend on the lengths of the query proteins A and B. However from the pseudocode we can also see that several other factors will affect the runtime: density of the interaction graph, length of the proteins in the graph ($|V|$ and $|X|$) and the size of the window (w). Table 4.1 below lists the cost in operations of each line and how many times each loop will be executed. For simplicity, we will ignore lines 1 and 2 and assume the query proteins and the interaction graph have already been loaded. We assume the FOR–loops have no cost except the lines contained within them. Therefore, the cost of a FOR–loop will be equal to its number of execution multiplied by the cost of the lines it contains.

Lines	**Costs**	**Loops Executed**
Line 3	-	$O(\|A\| - w)$
Line 4	-	$O(\|G\|)$
Line 5	-	$O(\|V\| - w)$
Line 6	$O(w)$	$O(1)$
Line 7	$O(\|R\|)$	$O(1)$
Line 8	-	$O(\|R\|)$
Line 9	-	$O(\|X\| - w)$
Line 10	-	$O(\|B\| - w)$
Line 11	$O(w)$	$O(1)$
Lines 12-13	$O(1)$	$O(1)$

Table 4.1: List of the PIPE algorithm pseudocode line costs and number of executions

Since the algorithm contains IF functions (lines 6 and 11) it will not always run every line of code found in the pseudocode. When the IF function in line 6 is always TRUE and the one in line 11 is always FALSE this causes the algorithm to do the maximum amount of work. In such a case the runtime of the PIPE algorithm is given by:

$$= O(|A| - w) \cdot O(|G|) \cdot O(|V| - w) \cdot (O(w) + O(|R|) + O(|R|) \cdot O(|X| - w) \cdot \\ \cdot O(|B| - w) \cdot (O(w) + O(1)))$$

The worst–case scenario for the PIPE algorithm consists of a fully connected interaction graph (every protein has every protein as its neighbor; $|R| = |G|$) and a window size of 1 ($w = 1$). In such a scenario the runtime becomes (we ignore w here since its value is 1 and is absorbed by the big-Oh notation):

$$= O(|A|) \cdot O(|G|) \cdot O(|V|) \cdot (O(|G|) + O(|G|) \cdot O(|X|) \cdot O(|B|))$$
$$= O(|A||G|^2|V|)(1 + O(|B||X|))$$

4.5 PIPE Parameter Tuning

There are three main parameters that need to be set for PIPE: (1) the window size w, (2) the threshold S_{pam} that determines a match between two fragments with respect to PAM120, and (3) the threshold M for the PIPE score (number of matches observed for two fragment a_i and b_j) above which PIPE reports an interaction between two proteins. The three values w, S_{pam} and M depend on each other. One of them can be set as a free parameter and the other two then need to be set accordingly. We chose to set the window size w to 20. Theoretically, one would want w to be as small as possible in order to identify interaction sites as precisely as possible. Although using a smaller window size would allow PIPE to detect small interaction sites (i.e. binding sites which are only up to 5-10AA long), too small a window size would create an overwhelming number of random matches. A window size of 20 is a small value for which the probability of random matches is small enough (see "Method 2" discussion below). We used two different methods to determine the values of the remaining two parameters, S_{pam} and M.

Method 1: Trial and error. For a set of 20 interacting pairs and 20 non–interacting protein pairs, we tried various combinations of S_{pam} and M, requiring close to 400 hours of computation time. It was observed that a PAM120 cut off score $S_{pam} = 35$ and a threshold for the number of matches $M = 10$ was most selective in differentiating between interacting and non–interacting pairs.

Method 2: Statistical evaluation. To evaluate the significance of $M = 10$ matches observed for a PAM120 cut off score $S_{pam} = 35$ with window size $w = 20$, we measured

the likelihood of such an event for random sequences. First, we built 1,000,000 random fragment pairs by creating 2,000,000 random fragments of length 20 whose amino acid distribution is the same as measured for our yeast database. Figure 4.2(a) shows the measured probability for two random fragments to match with a given PAM120 score (fragment score). We observe that the probability for two fragments to match with a PAM120 score larger than 35 is less than 10^{-6}. Next, we built 1,000 random protein pairs by creating 2,000 random proteins of length 500 whose amino acid distribution is the same as measured for our yeast database. For each protein pair, we ran PIPE and determined the maximum score in the PIPE result matrix. Figure 4.2(b) shows the measured probability for two random proteins to have a given maximum PIPE score. We observe that the probability of a PIPE score larger than 10 is less than 10^{-6}.

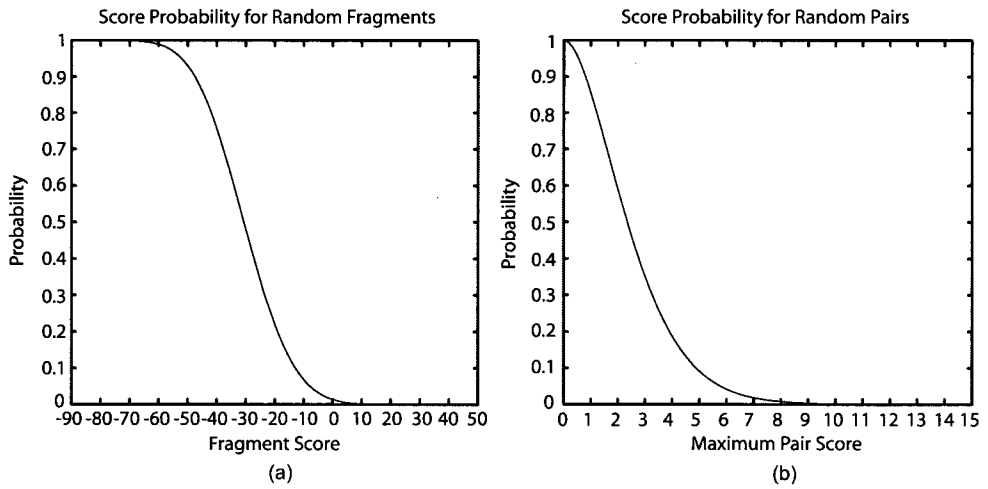

Figure 4.2: PIPE parameter tuning. In (a) the measured probability for two random fragments to match with a given PAM120 score (fragment score) is shown to be 10^{-6} for scores larger than 35. This was done using 1 M random fragment pairs of length 20. In (b) 1,000 random protein pairs of length 500 are used to show that the measured probability for two random proteins to have a maximum PIPE score larger than 10 is 10^{-6}.

4.6 Interpretation of PIPE output

Typical graphs of non–interacting and interacting pairs are shown in Figure 4.3(a) and Figure 4.3(b) respectively. The x and y axis represent the amino acids regions of the target proteins, starting from the N–terminal amino acid at position 1. Therefore position 5 corresponds to the 20 amino acid window starting at the fifth amino acid of the polypeptide. The score on the z axis represents the number of times that a pair of 20 amino acid sequences co–occurs in the dataset of interacting proteins. A high score corresponds to a high incidence of co–occurrence of the sequences among the database of interacting proteins. Therefore a score of 5 indicates that the corresponding sequences co–occur five times in our database, whereas a score of 50 indicates that the co–occurrence is present in 50 pairs of interacting proteins. We assume that a high score represents a soaring affinity for an interaction. We note that a major source of false positives reported by PIPE is motifs with frequent occurrence in the database. Pairs of such motifs can have a high co–occurrence simply because they are very frequent. PIPE's sensitivity and specificity is calculated as explained in Section 2.1.

4.7 Results

Our protein–protein interaction prediction algorithm (PIPE) relies on previously determined interactions for *S. cerevisiae*. For two target proteins A and B, PIPE determines the likelihood for A and B to interact. Typical PIPE output for non–interacting and interacting pairs of proteins are shown in Figure 4.3(a) and Figure 4.3(b) respectively. A peak with a score higher than 10 indicates that PIPE is predicting an interaction.

4.7.1 Ability of PIPE to detect interacting proteins

PIPE accuracy was determined by analyzing sets of known interacting pairs and expected non–interacting pairs. PIPE successfully detected 61% of interacting proteins in a randomly selected set of 100 protein pairs from the yeast protein interaction literature for which at least three different lines of experimental evidence supported the

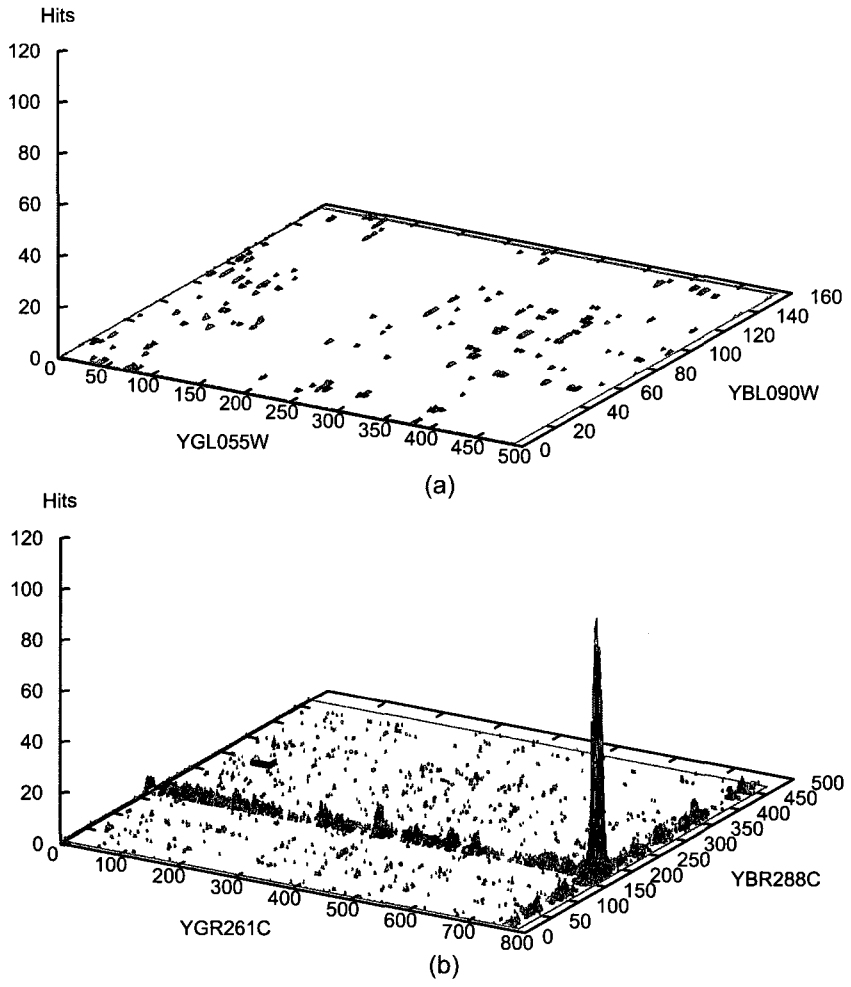

Figure 4.3: Interaction graphs. Two interaction graphs showing potential interaction sites for a pair of non-interacting proteins (a) and a pair of interacting proteins (b). In (a), the number of corresponding short amino acid sequences between YBL090W and YGL055W, which also co-occur in the dataset of the interacting proteins, is calculated to be very low and hence no obvious peaks are detected in this graph. In (b), a sharp peak with a score of 115 indicates that the two corresponding short amino acid sequences one in middle section of YBR288C (around amino acid 140) and the other at the C-terminal end of YGR261C co–occur 115 times in the dataset of the interacting protein pairs. It is therefore hypothesized that the two proteins YBR288C and YGR261C can potentially interact.

interaction. This positive validation set was selected independently of the dataset of the interacting protein pairs used by PIPE to predict interactions. This observation

suggests a sensitivity of 61% and a false negative rate of 39% for PIPE data. PIPE's success rate is comparable to those obtained by *in vivo* experiments. TAP tag data are estimated to have a false negative rate of 15–40% [29] with an internal reproducibility of 70% [28], which applies only to those proteins that can be successfully tagged *in vivo* (89%) [28]. A conservative estimation of false negative rate in yeast two–hybrid screens is approximately 40% [29, 114]. This finding indicates that protein interactions mediated by short polypeptide sequences may comprise the majority of protein interactions experimentally observed.

In order to evaluate the specificity and the rate of false positives associated with PIPE, a negative validation set of 100 protein pairs were gathered from the literature. These protein pairs are expected to not interact based on protein localization data, co–expression profiling, known direct or indirect functional or genetic relationships and the information gathered from the complete set of protein interaction datasets. 11 of these non–interacting protein pairs were predicted by PIPE to be interacting, indicating a specificity of 89% and a false/novel positive rate of 11%. It also suggests that PIPE has an overall accuracy of 75%. The low false positive rate associated with PIPE is substantially better than most experimental protein interaction detection methods. It is thought that the false/novel positive rate might be as high as 77% and 64% in TAP tag and yeast two hybrid experiments, respectively [29]. In addition to the negative validation set of 100 protein pairs discussed above, we also presented 10 pairs of random amino acid sequences of length 500 to PIPE, and PIPE detected no interactions among those 10 pairs, another indication of a low false/novel positive rate for PIPE (data not shown). All together these data indicate that PIPE can effectively identify protein–protein interactions based on the primary structure (amino acid sequences) of proteins alone and without any previous knowledge about the higher structure, domain composition, evolutionary conservation or the function of the target proteins. This is a significant improvement over some commonly used protein–protein interaction prediction algorithms. For example, our analysis using Interpret, one of the most commonly used protein–protein interaction prediction tools [5], failed to detect the previously identified interactions for protein pairs YKL028W–YDR311W, YKR048C–YCL024W [35, 127] and YOR358W–YGL237C [51] for which

limited structural information is available. PIPE analysis, however, detected high confidence interactions for these pairs with scores of 250, 160 and 100, respectively.

We note, however, that although PIPE appears to have a good specificity, it would be weak for detecting novel interactions among genome wide large–scale data sets. For example, assume that we were able to run PIPE on all pairs of yeast proteins (approx. 20,000,000 pairs), despite PIPE's current running time. If we assume that there are approximately 50,000 true interactions, then PIPE would be expected to report approximately 30,000 true positives, 2,200,000 false positives, 17,750,000 true negatives and 20,000 false negatives. The large number of false positives compared to the number of true positives makes PIPE a weak tool for analyzing such data sets. In Chapter 5 we will discuss improvements to the PIPE algorithm that will overcome these problems and enable genome–wide predictions.

4.7.2 Ability of PIPE to detect the sites of interactions between protein pairs

To examine whether PIPE can detect the sites of interaction between proteins, we took 10 protein pairs (Table 4.2) for which their sites of interactions had previously been reported. Of the 10 protein pairs, PIPE identified 7 pairs as interactors. The sites of

Protein A	Protein B
YPL153C	YBL051C
YNL088W	YGL017W
YNL243W	YBL007C
YCR084C	YBR112C
YMR190C	YNL282W
YGL153W	YDR244W
YBR079C	YNL243W
YDR477W	YGL115W
YMR159C	YMR159C
YDR216W	YMR303C

Table 4.2: Set of interacting proteins with previously reported interaction sites.

interactions reported by PIPE for 4 of these pairs were the same as those previously reported in the literature. It was previously shown in [41] that the region 310–768

in protein YNL243W is responsible for its interaction with amino acids 118-361 in protein YBL007C. PIPE analysis of the protein pair is shown in Figure 4.4. Apparent

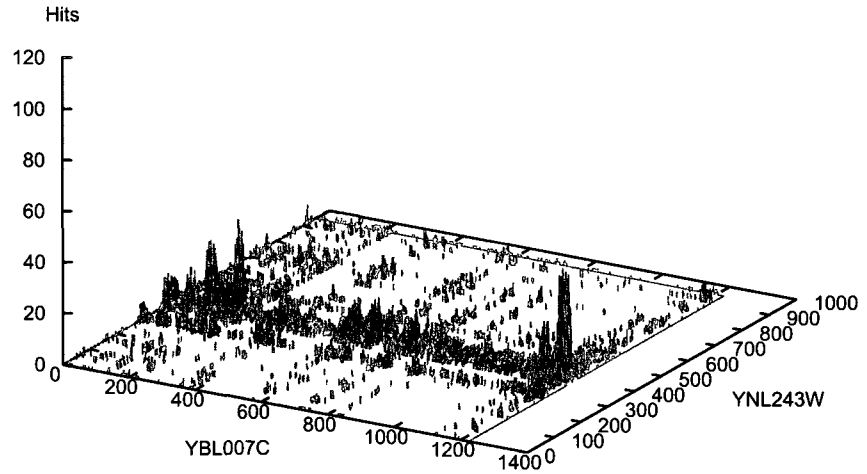

Figure 4.4: Potential interaction sites between YNL243W and YBL007C. PIPE can successfully determine the previously known sites of interaction between the two proteins YNL243W and YBL007C. It was previously shown that the region 310–768 in protein YNL243W is responsible for its interaction with amino acids 118–361 in protein YBL007C. Visualized by its highest peak with a score of 45, PIPE has successfully detected an interaction between YNL243W and YBL007C via their corresponding amino acid regions 350–410 and 100–250, respectively. A second highest peak with a score of 42 also suggest a second site of interaction between the two proteins. According to PIPE analysis it is possible that the C-terminal end of the YBL007C protein may also serve as a site of interaction.

by a peak with a high score of 45, PIPE analysis indicates that the region between amino acids 350 and 410 in protein YNL243W co-occurs frequently with the region between amino acids 100 and 250 in protein YBL007C. This observation suggests that the two proteins are interacting via the mentioned regions. This is in agreement with the regions experimentally shown to mediate an interaction between YNL243W and YBL007C [41]. Interestingly, PIPE also detected a second potential site of interaction between the same region (amino acids 350–410) for YNL243W as above and the C–terminal region (amino acids 1175–1225) of YBL007C. Of interest is that previously it was shown that the C-terminal domain of YBL007C can function as a site of protein–protein interaction [121, 140]. Further studies are required, however, to verify

the presence of an interaction between these newly predicted sites. Furthermore, PIPE successfully determined the previously documented site of interaction between YCR084C and YBR112C. It is reported that the first 75 amino acids of YCR084C is responsible for an interaction with the N–terminal region of YBR112C [125, 126]. PIPE correctly predicted an interaction between these two sites. In addition PIPE analysis successfully predicted the known interaction site between YBR079C and YNL243W [86] as well as the region responsible for dimerization of YMR159C [81]. All together, this data indicates a 40% success for PIPE to identify the previously reported interaction sites between proteins. We note that this success rate is measured from a very small data set since there is not much reliable data available that correctly identifies the sites of protein interactions.

4.7.3 Ability of PIPE to detect novel protein–protein interactions

The ability of PIPE to detect novel protein–protein interactions was examined by analyzing the potential interaction between a novel pair of proteins, YGL227W–YMR135C for which no experimental interaction data was available when we initiated this project. Little is known about the molecular function of these genes, but the inactivation of either YGL227W or YMR135C, also known as VID30 and GID8, respectively, are shown to alter proteasome dependent catabolite degradation of fructose–1,6–bisphosphatase (FBPase) [100]. PIPE analysis of this protein pair is shown in Figure 4.5(a). The peak score of 136 indicates that the proteins are capable of interacting with one another. This is in agreement with the phenotypic characteristics of the yeast strains in which either YGL227W or YMR135C is deleted. Both deletion strains are incapable of degrading FBPase [100]. To confirm the validity of the observed interaction, TAP tag methodology was employed. An advantage of TAP–tagging over other generic protein–protein interaction detection assays is that it detects those interactions that occur under native level of protein expression in the cell. Therefore, TAP tag identifies those complexes that really exist *in vivo*. As shown in Figure 4.5(b) when YGL227W is TAP–tagged and its corresponding complex is affinity purified, YMR135C is identified as an interacting protein partner. The LC–MS MS analysis also indicated that YMR135C co–purified as an interacting partner

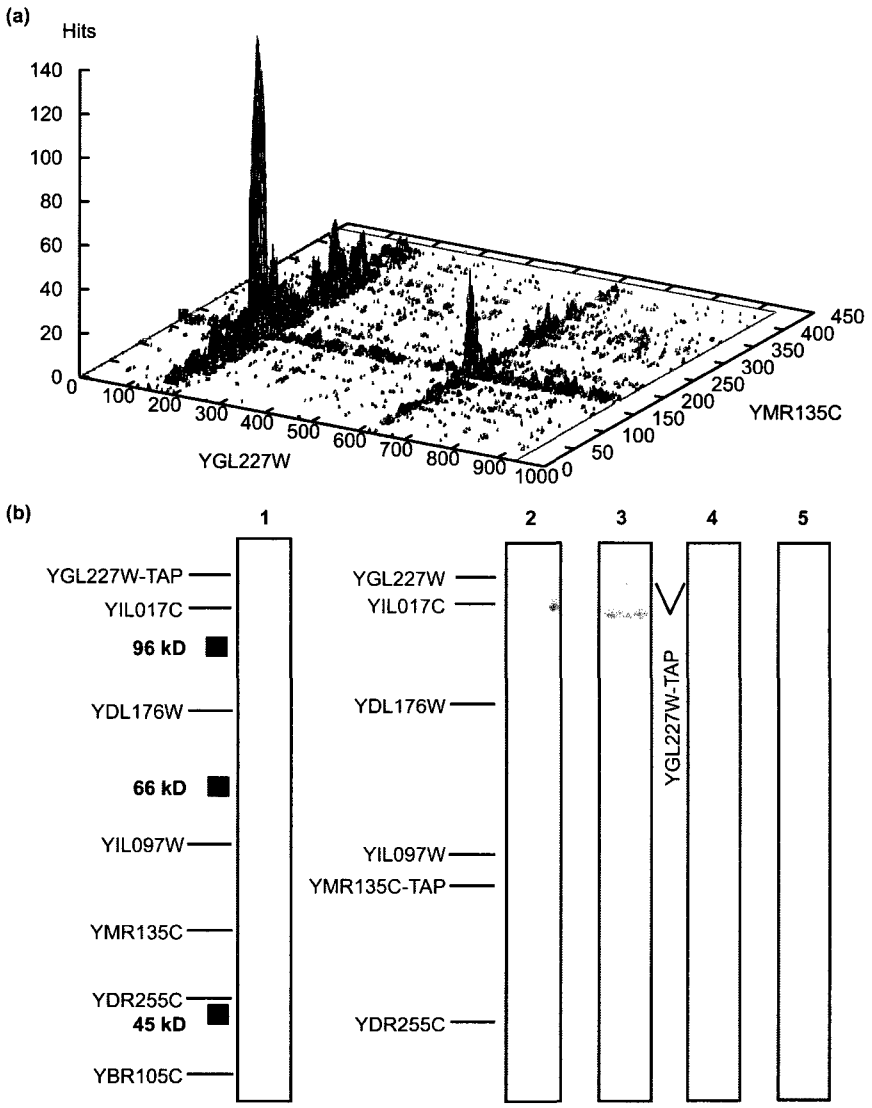

Figure 4.5: Novel protein–protein interaction identified by PIPE. With a score of 136 in (a), PIPE analysis predicts an interaction between YGL227W and YMR135C. (b) TAP-tag analysis confirms the interaction between YGL227W and YMR135C. When YGL227W is TAP-tagged, YMR135C is purified as an interacting subunit (panel 1). Reciprocal TAP-tagging of YMR135C also identifies YGL227W as an interacting partner (panel 2). Panels 3 and 4 show the purifications of TAP-tagged YGL227W strains in which either YDR255C (YDR255CΔ) or YMR135C (YMR135CΔ) were deleted, respectively. Deletion of YDR255C (panel 3) had no effect in the co-purification of other subunits. However, when YMR135C (panel 4) was deleted, the interactions between TAP-tagged YGL227W and most other subunits were eliminated. Panel 5 is used as a control and shows the purification of a strain, which is not tagged.

when TAP–tagged YGL227W was purified. The reciprocal tagging and purification of YMR135C confirmed this interaction. YGL227W was identified as an interacting partner when TAP-tagged YMR135C complex was affinity purified. The presence of YGL227W in the purified mixture was also verified by LC–MS MS analysis. All together, these data demonstrate that PIPE has the ability to successfully predict novel protein–protein interactions.

Besides the obvious advantages of PIPE over TAP tagging (speed and the ease of use), PIPE can also be used to analyze yeast proteins for which TAP tagging fails. A genome-wide yeast TAP tagging project has indicated that out of the 6,466 yeast open reading frames, only 1,993 (or 31%) can be successfully TAP–tagged and purified [36]. Data from the same authors [36] suggest that TAP tagging of YCR093W was unsuccessful. However, with a score of 60, PIPE analysis successfully identified a previously known interaction between YCR093W and YPR072W [84]. Since the screening of yeast complexes to saturation using TAP tag has identified approximately 62% of the expected yeast protein complexes [36], it might be expected that a different approach like PIPE may be able to contribute to the identification of some remaining interactions.

4.7.4 Ability of PIPE to elucidate the internal architecture of protein complexes

TAP tagging of YGL227W resulted in the co–purification of six other proteins (YIL017C, YMR135C, YDL176W, YIL097W, YBR105C and YDR255C) as indicated in Figure 4.5(b). This suggests that YGL227W forms a novel protein complex with these proteins that here we term vid30 complex (vid30c). The presence of this protein complex is further confirmed by TAP tagging of YMR135C, which resulted in the co–purification of the same constituent subunits; see Figure 4.5(b). The internal architecture of this protein complex, however, remains unknown, as TAP tag has a limited ability to resolve the internal structure of complexes. To test the ability of PIPE to provide a better understanding of the internal architecture of protein complexes, we systematically analyzed protein pairs of vid30c constituent subunits using

PIPE. This resulted in the analysis of 21 protein pairs, the result of which is summarized in Table 4.3. This data was then used to generate a hypothetical representation

Protein A	Protein B	Score
YBR105C	YDL176W	5
YBR105C	YDR255C	4
YBR105C	YGL227W	27
YBR105C	YIL017C	23
YBR105C	YIL097W	4
YBR105C	YMR135C	10
YDL176W	YGL227W	75
YDL176W	YIL017C	46
YDR255C	YDL176W	4
YDR255C	YGL227W	26
YDR255C	YIL017C	17
YDR255C	YIL097W	4
YDR255C	YMR135C	5
YIL017C	YGL227W	460
YIL097W	YDL176W	6
YIL097W	YGL227W	50
YIL097W	YIL017C	29
YMR135C	YDL176W	21
YMR135C	YGL227W	136
YMR135C	YIL017C	105
YMR135C	YIL097W	9

Table 4.3: Internal PIPE scores for vid30c. PIPE scores are used to show the potential interactions between the subunits of vid30c.

of how the protein subunits might be interacting. As shown in Figure 4.6, vid30c seem to have a core component consisting of four subunits YGL227W, YIL017C, YMR135C and YDL176W. These four subunits seem to be in direct interaction with each other. The complex also seems to have a secondary component, the members of which (YIL097W, YBR105C and YDR255C) seem to interact with YGL227W and YIL017C only and not to each other. The hypothesized interactions among the subunits of the core component seem to have high PIPE scores suggesting high affinity and likelihood for interactions. The PIPE scores associated with the secondary components, however, tend to be lower. The highest PIPE score (460) was that for the interaction between YIL017C and YGL227W, which might be expected, as all the

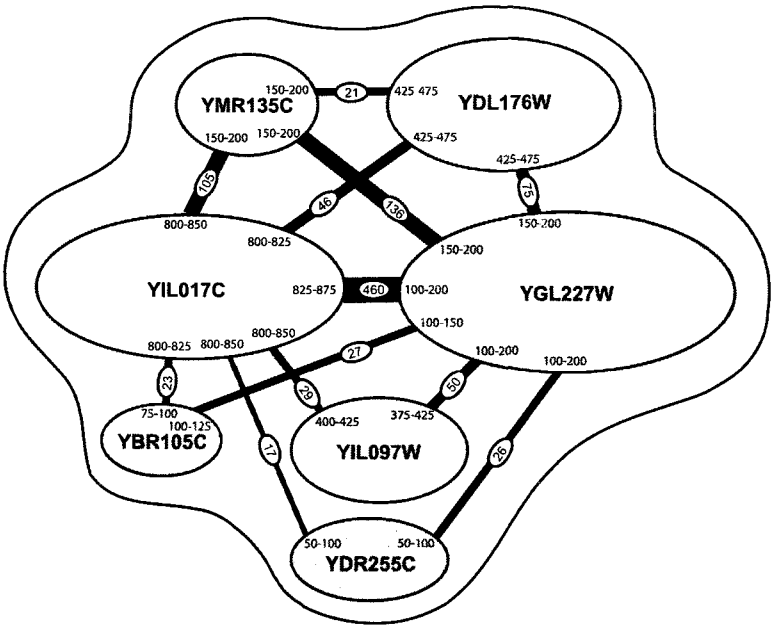

Figure 4.6: Internal architecture of vid30c as suggested by PIPE. YGL227W, YIL017C, YMR135C and YDL176W all interact with each other with relatively high PIPE scores, and seem to form a core compartment of vid30c. YIL097W, YBR105C and YDR255C, with relatively lower PIPE scores, interact with YGL227W and YIL017C only, and not with each other, suggesting a secondary component of vid30c. PIPE scores are embedded within the connecting lines. The regions responsible for interactions are indicated.

subunits of vid30c seem to interact with these two proteins. The lowest significant PIPE score was for YDR255C, which only had two significant scores, 25 and 17, for interactions with YGL227W and YIL017C, respectively, suggesting a low affinity (implied by a low PIPE score) for an interaction with vid30c. The hypothetical sites of interactions identified by PIPE are different in size. For example, YIL017C seem to interact with a small region of YBR105C (75–100), and with a relatively broader region of YGL227W (100–200). It also seems that each protein may have a specific region responsible for interaction with protein partners. This in turn may suggest that some of these proteins may compete for an interaction with the same partner. There remains the possibility however, that the broader regions (such as YGL227W region 100–200) may support simultaneous interactions with more than one protein

partners.

To experimentally examine the information from PIPE analysis about the internal topology of vid30c, we made two gene deletion strains. For this purpose YDR255C and YMR135C were selected which have similar molecular weights (50 and 52 kD, respectively). According to PIPE, YDR255C has the lowest affinity to vid30c. Therefore, it might be expected that the deletion of this gene may be insignificant to the integrity of vid30c. However, PIPE analysis placed YMR135C in the core component of vid30c. Depending on the molecular function of YMR135C, it might be expected that the elimination of this protein may (or may not) alter the formation of vid30c. Therefore, two yeast deletion strains, YDR255CΔ and YMR135CΔ, were generated in which either the YDR255C or YMR135C gene was deleted, respectively, in a TAP–tagged YGL227W yeast background. In agreement with PIPE analysis, TAP tagging of YDR255CΔ strain indicated that deletion of YDR255C showed no significant effect in the formation of vid30c; see Figure 4.5(b). Besides YDR255C all other members of vid30c co–purified with TAP–tagged YGL227W. However, when YMR135C was deleted (YMR135CΔ), the interactions between TAP tagged YGL227W and most other vid30c subunits were eliminated; see Figure 4.5(b). This suggests that vid30c was not formed in the absence of YMR135C. This is in agreement with PIPE analysis, which indicated a low affinity between YDR255C and vid30c, but placed YMR135C in the core component of vid30c with strong affinity to this complex.

To estimate the success rate of PIPE in predicting the internal structure of protein complexes, we tested PIPE on 10 protein complexes (see Table 4.4). Each complex consists of three subunits, and the subunits are reported to be interacting with each other in a chain format, that is "*A*–*B*–*C*", where protein A interacts with B but not with C, and protein C interacts with B only. It should be noted however, that due to the technical limitations associated with the approaches used to generate our current view of the internal structure of protein complexes and in the absence of a sufficient number of studies on the crystal structural analysis of protein complexes, the topology of the reported complexes should be considered with caution. Regardless, these 10 protein complexes generated a total of 30 potential interactions, 20 of which

were shown to exist in interaction databases and 10 of which were shown not to. PIPE detected 13 interactions of the 20 shown to exist. It also detected 4 false/novel interactions of the 10 shown not to exist. In total, from the 10 protein complexes, PIPE detected 3 internal architectures identical to what was reported previously. It should be noted that due to the absence of more reliable data, this may not represent the true success rate of PIPE but instead represents the overlap between the existing small data set and the data generated by PIPE.

Protein A	Protein B	Protein C
YGR004W	YLR324W	YDL089W
YPR119W	YBR135W	YDL155W
YDR378C	YER112W	YJL124C
YMR197C	YOR036W	YBL050W
YPR185W	YGR120C	YPR105C
YKL103C	YAL034W	YGR120C
YBL026W	YDL160C	YEL015W
YLR423C	YGR113W	YIL144W
YDR084C	YGL198W	YPL095C
YLR045C	YCL029C	YER016W

Table 4.4: Set of protein complexes with previously reported internal structures. This list was used to evaluate the efficiency of PIPE to predict the internal architecture of protein complexes. Only the adjacent subunits are reported to be interacting.

4.7.5 Discussion of the algorithmic approach

The PIPE method predicts the likelihood of interaction between two query proteins A and B by measuring how often pairs of subsequences in A and B co–occur in pairs of protein sequences in the dataset that are known to interact. The amount of computation involved is substantial. For a pair of interacting proteins, on average, several hours of computation time were required for a standard desktop machine. This time was observed to be directly proportional to the number of re–occurrences of similar sequences in different interacting proteins in our dataset of interacting protein partners. As the number of corresponding sequences that co–occurred in the dataset increased, so did the computation time associated with analyzing the target protein pair. Similarly, the computation time required for non–interacting

protein pairs were observed to be significantly lower as the co–occurring sequences were absent in these pairs. In Chapter 5, we will discuss how considerable speed improvements were implemented to reduce this runtime. The original version of PIPE concentrates on the predictive precision of the method but we will discuss in Chapter 5 the process of applying more sophisticated data structures and algorithms to reduce PIPE's computation time. In addition, PIPE has been parallelized so that it can be executed on a processor cluster instead of a single workstation.

4.8 Summary

In this chapter we discussed the making of a computational tool, termed PIPE, which can effectively identify protein interactions among *S. cerevisiae* protein pairs. The sensitivity of this engine to identify true interactions is estimated to be 61%, which is comparable to that of the currently available generic biochemical assays used for large–scale detection of protein–protein interactions. PIPE has an estimated specificity of 89%, which is a significant improvement over the currently available confidence rates for most other assays. In addition, PIPE considerably reduces the cost associated with detecting protein interactions by traditional biochemical methods. In Chapter 5, we will discuss how considerable speed improvements were implemented to reduce this runtime. Furthermore, by incorporating additional protein interaction data into PIPE's database, as well as using more precise tools for detecting similar short polypeptide sequences in different proteins (e.g. allowing for gaps), we hope to further increase the precision of PIPE in the future. In addition, the incorporation of the data gathered from three–dimensional structures of proteins and protein complexes is also expected to further enhance the ability of PIPE to detect protein–protein interactions. The fact that protein–protein interactions can be successfully detected from the amino acid sequences of proteins alone and without additional information/predictions about the proteins has also led to the development of another version of PIPE for predicting interactions in other organisms. This version will be discussed in Chapter 6.

Chapter 5

PIPE2: Improving PIPE

5.1 Introduction

In the previous chapter we've introduced the PIPE algorithm and proven that it can be used to successfully detect protein–protein interactions in *S. cerevisiae* (budding yeast). However we also saw there were major obstacles to overcome before we can run all possible protein pairs in yeast or any other organism. Although PIPE is successful in identifying novel interactions, two issues preclude it from being used in a proteome–wide investigation to discover potential PPIs: i) it is computationally expensive requiring hours of computation per protein pair and ii) the 89% specificity would generate a tremendous number of false positives if applied to all possible protein pairs in a proteome.

In this chapter we describe our efforts to systematically investigate all potential yeast protein interaction pairs using an improved sequence–based computational method and develop a more comprehensive PPI map for yeast. With the creation of the second version of PIPE, called PIPE2 [95], we significantly increase the speed of execution and the specificity of PIPE. Analysis of the resulting network yields new insights into the protein interaction network and provides new avenues for investigating potential interactions not amenable with recent large–scale experimental investigations. An online PIPE2 portal has been made available at http://pipe.cgmlab.org/ along with executable binaries and our complete dataset. A paper published by Pitre et al. [95] is partly the basis for this chapter.

5.2 The Problem

Due to the brute force exhaustive scan of the proteins and database in the PIPE algorithm, running a prediction on a single pair could potentially take hours of runtime

on a single processor. It is easy to see that a genome-wide scan of yeast which contains roughly 6400 proteins and therefore a possible 20 million pairs could take millions of hours on a single processor (several years using hundreds of processors). The original PIPE algorithm's runtime needs to be improved by several orders of magnitude to make this experiment possible.

The second issue with a genome-wide scan using the original PIPE algorithm is the number of expected false positives. We've seen that the specificity of PIPE has been estimated at 89%. While this value puts it above the traditional methods currently in use it still is too low to scan a large number of protein pairs. For example using 100 pairs we could expect 11 false positives which could be acceptable depending on our requirements. However scanning 20 million pairs we could expect 2,200,000 false positives which is unacceptable when we take in account the predicted number of interactions in yeast varies between 16,000 and 26,000 [42]. Even a 99.0% specificity (1% false positives) would yield 200,000 false positives which is 8–12x larger than the expected number of true interactions. We can usually increase the specificity at the expense of lowering the sensitivity. An acceptable solution would increase the specificity so that the number of false positives is below the expected number of interaction while keeping the sensitivity high enough to still detect interactions.

5.3 Time Complexity Improvement

The speed improvements described in this section were done by Chris North as part of his M.C.S. thesis at Carleton University (greater details can be found there).

5.3.1 Optimizing Window Comparisons

Potential areas for improvement were obtained by function level profiling using the GNU profiling tool gprof. gprof uses time–domain sampling to estimate the amount of time spent in each function. It also implements automatic program instrumentation to determine the number of times each function is called and to record the call graph information during program execution. A profile of the original PIPE implementation indicated that over 99% of the runtime was spent performing the fixed–length window comparisons. This realization led to two immediate optimizations resulting

in a considerable performance improvement.

The first optimization converted the character–based amino acid representation to a binary representation. This simple change removed the need for a character–to–index mapping function when applying the PAM120 similarity matrix. The optimization resulted in an 18x performance improvement when performing the window comparison operation.

The second optimization involved implementing a sliding window approach for the window comparisons. This approach takes advantage of the amino acid sequence overlap in consecutive windows, thus reducing the number of PAM120 lookups required. A similar optimization is used in the Paircomp comparative analysis tool [17]. This optimization resulted in a 10x performance improvement when performing the window comparison operation in PIPE.

5.3.2 Pre–Computation & Query Approach

Despite the performance improvements achieved by the two window comparison optimizations, the large scale investigation of all possible yeast protein pairs remained computationally infeasible. Other approaches to reducing the computational cost of PIPE were therefore required.

One approach that proved fruitful was to eliminate the repetition of the same window comparisons throughout the evaluation of all possible protein pairs. In the original PIPE algorithm, each protein pair is evaluated independently and there is no mechanism for pre–computing or caching of sequence window comparisons between query proteins and library proteins. By implementing such a mechanism it was believed that the overall runtime would improve due to a reduction in the number of repeated window comparisons at the expense of additional memory required for storing the previously computed comparisons. The implementation of this approach required two stages: the pre–computation stage and the query stage. During the pre–computation stage, all possible query protein windows are compared with all sequence windows from all library protein sequences. Only those window pairs that are deemed to be similar are recorded. Once this initial one–time work is completed,

subsequent PIPE2 runs can use these results without re–computing the window comparisons. The results of these comparisons are written to non–volatile storage media (e.g. local disk). During the query stage when a query protein pair is evaluated for possible PPI, the pre–computed window similarity information is loaded into main memory and can be rapidly queried as required.

5.3.3 PIPE2 Pseudocode

The PIPE2 algorithm is essentially the same as described in Chapter 4; the results are exactly the same but what changes is the amount of work being done. Instead of doing all the comparisons "on the fly", matches are pre–computed and stored on disk locally. Here we discuss the algorithmic differences and what the impact will be on the complexity. **procedure** *PIPE2*(G, H, A, B)

1: Input graph G
2: Input query proteins A of length m and B of length n
3: **for** every a_i in A of length w for $i = 0$ to $(|A| - w)$ **do**
4: load list of proteins that match a_i
5: **for** every protein V that match a_i **do**
6: create a list R containing every neighbor of V in G
7: **for** every protein X in R **do**
8: **for** every b_j in B of length w for $j = 0$ to $(|B| - w)$ **do**
9: load list of proteins that match b_j
10: **if** list of proteins that match b_j contains X **then**
11: $H[i][j] = H[i][j]+1$
12: stop comparing again X and go to next neighbor in R
13: **end if**
14: **end for**
15: **end for**
16: **end for**
17: **end for**
18: **return** H

From the PIPE2 pseudocode we notice a lot of FOR loops have been removed. We no longer need to scan the entire protein list for a match for each a_i since we have that list pre–computed and can simply be loaded instead. We also don't need to compare every window from protein B against every neighbor X in R since we can simply check if X appears in the list of proteins that match the correct window in B. Let us assume the worst–case that every protein in the graph G will match windows a_i and b_i so that the lists we load in lines 4 and 9 will be of size $|G|$. Table 5.1 below lists the cost in operations of each line and how many times each loop will be executed. As with the PIPE pseudocode we will ignore lines 1 and 2 and assume the query proteins and the interaction graph have already been loaded. We also assume the FOR–loops have not cost except the lines contained within them.

Lines	**Costs**	**Loops Executed**
Line 3	-	$O(\|A\| - w)$
Line 4	$O(\|G\|)$	-
Line 5	-	$O(\|G\|)$
Line 6	$O(\|G\|)$	-
Line 7	$O(\|G\|)$	-
Line 8	-	$O(\|B\| - w)$
Line 9	$O(\|G\|)$	-
Line 10-12	$O(1)$	-

Table 5.1: List of the PIPE algorithm pseudocode line costs and number of executions

If we sum each line, the total complexity for the PIPE2 algorithm is:

$$= O(|A| - w) \cdot O(|G|) \cdot (O(|R|) + O(|R|) \cdot O(|B| - w) \cdot O(|G|))$$

And if we assume again that $w = 1$ we get:

$$\begin{aligned} &= O(|A|) \cdot O(|G|) \cdot (O(|R|) + O(|R|) \cdot O(|B|) \cdot O(|G|)) \\ &= O(|A||G||R|) + O(|A||B||G|^2|R|) \end{aligned}$$

We can see already that compared to the original PIPE complexity, PIPE2 is no longer dependant on the length of the proteins in the database ($|V|$ and $|X|$ in the PIPE pseudocode). The number of proteins in the graph ($|G|$) becomes an important

factor in the complexity but that is simply because we assumed a fully connected graph. In reality, interaction graphs are VERY sparse and the number of neighbors per protein would be nowhere close to $|G|$. The fact that we don't have to do any window comparisons here (*compare* function in the PIPE pseudocode) explains the huge speedup discussed in the *Results* section of this chapter.

5.4 Increased Sensitivity and Specificity

In the original PIPE algorithm [94] a positive interaction was identified by applying a threshold test on the window comparison result matrix H. If the maximum value of H was above the minimum threshold value of 10, then query proteins A and B were deemed to interact. This yielded 61% sensitivity and 89% specificity over a relatively small test set of 100 positive and 100 negative interactions. However, if such a false positive rate of 11% were applied over 20 million possible pairs in an all–to–all experiment, we would expect to generate 2,200,000 false positives. Given that the expected number of total interactions in *S. cerevisiae* is estimated to lie between 16,000 and 26,000 [42] it was necessary to increase the specificity to over 99.9%, thereby reducing the false positive rate to less than 0.1%. To this end, several threshold values were investigated to increase the specificity and reduce the false positive rate. Through a manual inspection of several result matrix 'landscapes', it was noted that true interactions often resulted in hill–like regions with broad support, while false positives were characterized as thin lines of high scoring pairs, which often reflected a local region of low complexity. In contrast to our previous approach of applying a moving average filter, we applied a median filter, which effectively eliminated thin line regions and maintained hill regions. For a cell c in the matrix, a median filter evaluates the surrounding $f \times f$ values f being the width of the filter and always being odd. Those f^2 values are then sorted and the cell c is replaced by the median value (see Figure 5.1).

Unfortunately, the median filter is computationally expensive – a naïve implementation of the median filter initially took longer than the entire PIPE2 algorithm for large filter sizes. Since the median filter relies on sorting the f^2 within the filter area we can bound the running time on an n by m matrix by $mnf^2 \log_2 f^2$ for a filter of

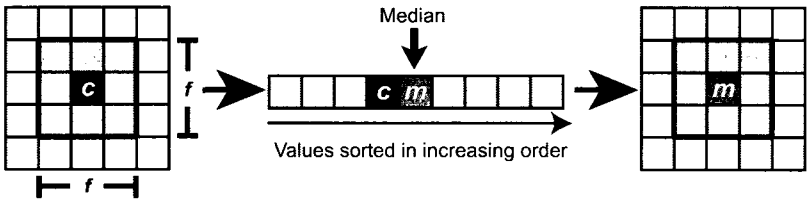

Figure 5.1: A median filter assigns the median value of an fxf matrix centered by cell c.

width f. Of course this upper bound assumes we must sort the f^2 numbers from scratch at every step. Instead we can take advantage of the fact that we can move the filter by one cell at every step. In Figure 5.2(b) and (c) we can visualize what happens when a 3 × 3 filter moves right by one position in the matrix: three new cells are added to the filter area (rightmost column in the filter area) and three are removed (leftmost column in the filter area). By moving the filter in a snake–like

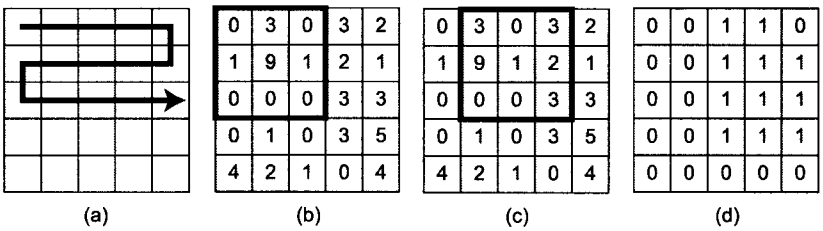

Figure 5.2: An example of running a 3 × 3 binary median filter on a 5 × 5 matrix. (a) Movement of the filter in the matrix. (b) Filter area for cell (1,1), the center cell being the value modified in this step. (c) The resulting matrix after applying the binary median filter to the matrix shown in (b).

fashion (5.2(a)) we see that at every move of the filter (right, left, down or up) this will always remain true. Once the first filter area has been sorted (for matrix position (0,0)) it can be updated at every step by performing three deletions and three insertions. With this slight improvement we can reduce the runtime of the median filter to: $f^2 \log_2 f^2 + 2f^2mn$ ($f^2 \log_2 f^2$ to sort the first filter area, then for every filter move we must delete f numbers from the filter area and insert f numbers, each insert/delete taking at most f operations). For a 3 × 3 filter and two query proteins of length 500AA ($m = 500, n = 500$) this reduces the number of operations from

7.1M using the naïve approach to 4.5M using the snake–like method (1.58x less operations). Increasing the filter size will of course yield even better improvements (by $\approx \log_2(f^2)/2$ less operations). This runtime was still considered too high since for larger proteins ($>$ 500AA) the filter would often take longer to run than the PIPE2 prediction algorithm.

In order to reduce running time further, the filter was modified such that if there are more zero neighbors of c than non–zeros, c is changed to 0, otherwise c is changed to 1. This eliminates sorting and speeds up the filtering, particularly with large filter sizes. This counting method reduces the runtime of the filter on an $m \times n$ matrix to: $f^2 + 2fmn$. In our previous example with two query protein of length 500AA and a 3×3 filter, this simplified binary median filter further reduces the number of operations from 4.5M to 1.1M.

We can calculate the speedup over the naïve median filter implementation as:

$$\begin{aligned} &= \frac{f^2 mn \log_2 f^2}{f^2 + 2fmn} \\ &= \frac{O(2f^2 mn \log_2 f)}{O(2fmn)} \\ &= O(f \log_2 f). \end{aligned}$$

We can also calculate the speedup over the snake–like median filter implementation as:

$$\begin{aligned} &= \frac{f^2 \log_2 f^2 + 2f^2 mn}{f^2 + 2fmn} \\ &= \frac{O(2f^2 mn)}{O(2fmn)} \\ &= O(f). \end{aligned}$$

We can see that in general, this approach yields a reduction of operations by a factor of $O(f \log_2 f)$ compared to the naïve median filter implementation or by $O(f)$ compared to the snake–like median filter.

Following the application of the binary median filter, the average value of H is then compared to a threshold value to determine whether the proteins are interacting. The filter size and threshold value were optimized via leave–one–out cross–validation

(LOOCV) tests over a true positive dataset and a true negative dataset (Figure 5.3). Each experiment consisted of running the positive and negative datasets through PIPE2 while varying the filter size as well as the average cutoff while measuring the sensitivity and specificity. Traditionally, AUC (Area Under the Curve) is used to quantify and compare different ROC curves. Although the AUC would indicated how well the algorithm is doing in general for all cutoffs, the large number of expected false positives at low specificities leads us to only concentrate on a small section of the curve (typically > 90% specificity). Therefore the AUC would not be indicative of the performance we are looking for at high-specificities and will not be shown for the ROC curves. The true positive dataset was generated by taking the intersection of reported

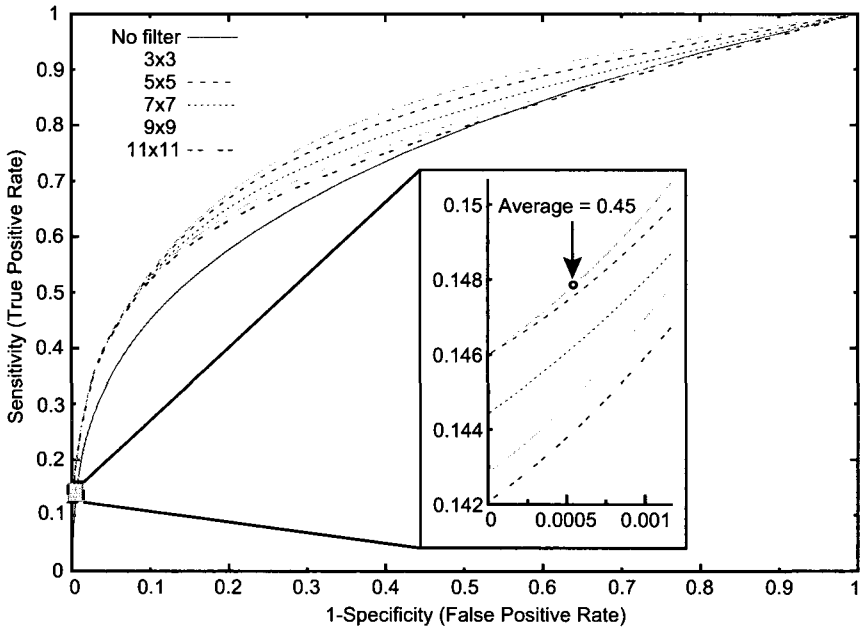

Figure 5.3: Receiver Operating Curve (ROC) by varying filter size and average cutoff. A binary median filter size of 3x3 with a threshold of 0.45 provides optimal sensitivity to achieve a false positive rate of 0.001. *Insert:* A closer look at the curves for high specificities (> 99.9%).

PPIs from BioGRID [119], Krogan *et al.* [61] and our original dataset for PIPE [94] which yielded 1,274 interactions. Since these PPIs were independently reported in at least three studies, these were considered to be 'true positive' interactions for the

purpose of evaluating the sensitivity of PIPE2. In order to measure the specificity of PIPE2, we also needed a negative dataset of approximately the same size but since a large negative database does not exist it had to be created. A true negative set of 1274 pairs was constructed by choosing two proteins at random [13] then checking to make sure that the candidate negative pair is not reported in either our database or in BioGRID [119]. During LOOCV testing, each of the 1274 positive pairs from the true positive dataset is removed individually from our PPI library prior to running that pair through PIPE2. This prevents our algorithm from identifying a pair as positive simply because it is already present in our database. Two types of experiments were performed: "No filter" and "Filter and Cutoff". For the "No filter" experiment the average cutoff threshold was varied from 0.0 to 1.5 in 0.01 increments. In the "Filter and Cutoff" experiments the filter was varied from 3x3 to 11x11 and the average threshold from 0.0 to 1.0 in 0.01 increments. The results of this experiment are illustrated in Figure 5.3. For these experiments, the desired specificity was fixed at 99.9% and the sensitivity to true positive interactions was optimized through filter design and selection of average threshold. It can be seen from Figure 5.3 that the combination of the median filter and application of a cutoff on the average value is important to achieve reasonable sensitivity rates when we set our goal specificity to 99.9%. A filter size of 3x3 and average cutoff of 0.45 achieves a sensitivity of 14.6% while meeting our design constraint of a specificity of 99.9% over the test set (accuracy of 88.7%). Increasing the filter size beyond this point unduly lowers the sensitivity, while lowering the average cutoff lowers the specificity below 99.9%. As reported we have seen in Chapter 4, actual sites of interaction may often be determined through examination of the window comparison output matrix H. When this matrix is visualized as a landscape, the actual interaction sites may appear as a steep hill or peak. Note that for high–throughput all–to–all applications of PIPE2, the matrix H is not reported due to resource constraints. However, for more detailed analysis of individual protein pairs, the raw matrix (i.e. prior to application of the binary median filter) is made available to the user.

5.5 Results

5.5.1 Performance Speedup

PIPE2 incorporates two window comparison optimizations and one structural change over the initial algorithm. Table 5.2 shows each change along with the average single–processor runtime per PPI prediction and the overall performance improvement over the original PIPE implementation. These runtime numbers were obtained after running the program on the same set of 1,000 randomly chosen protein pairs.

The PIPE2 implementation resulted in a 16,150x performance improvement over the original algorithm and implementation. This made it possible to run PIPE2 against all 20 million protein pairs in *S. cerevisiae* in approximately two days using a 76–processor Linux Cluster rather than the estimated 50 years that would have been required using the original algorithm. The tremendous performance improvement credited to the Pre–computation / Query Approach is not entirely accurate since it does not take into account the time spent performing the pre–computation of all possible window comparisons. However, this pre–computation time (30 minutes) is only a small percentage (1%) of the total all–to–all runtime (48 hours on 76 processors), so the improvements provided by this new approach are still significant (14,775x). The motivation for the Pre–computation / Query Approach was to eliminate the repetition of the same window comparison throughout the evaluation of all possible protein pairs. In order to explain the speedup achieved, counters were added to the original PIPE implementation and the fully optimized PIPE2 implementation (pre–computation stage only) to keep track of the total number of window comparisons performed. When both PIPE and PIPE2 were applied to a set of 1,000 randomly chosen protein pairs and then extrapolated to the entire all–to–all experiment for

Version	Average Runtime	Speedup
Original PIPE	6,944.40 sec.	1x
+ Digital Alphabet Optimization	389.65 sec.	18x
+ Sliding Windows Optimization	160.53 sec.	43x
+ Pre–Computation / Query Approach (PIPE2)	0.43 sec.	16,150x

Table 5.2: Successive performance improvement of PIPE to PIPE2. Using all improvements PIPE2 now runs more than 16,000 times faster than the original PIPE

S. cerevisiae, PIPE2 was estimated to perform 25,000 times fewer window comparisons over the original PIPE algorithm, which explains the tremendous performance improvement provided by the Pre–Computation / Query Approach.

5.5.2 Evaluation of sensitivity and specificity

In *Section 5.4:Increased Sensitivity and Specificity*, we tuned our parameters by using a true positive set and a true negative set of 1274 pairs each. In order to better evaluate the specificity of PIPE2, a larger set of true negatives was needed. Therefore, we constructed a negative set of 100,000 randomly chosen pairs as explained in [13] that are not reported in either our database or in BioGRID [119]. To evaluate the sensitivity of PIPE2, we used the true positive dataset of 1,274 interactions by taking the intersection of reported PPIs from bioGRID [119], Krogan *et al.* (S7 set) [61] and our original dataset for PIPE [94]. For the true positive dataset, PIPE2 correctly identified 186 pairs as true positives (1,088 false negative), which results in a sensitivity of 14.6%. For the true negative dataset PIPE2 correctly identified 99,946 as true negatives (54 false positives) yielding a 99.95% specificity. It should be noted that such a randomly selected negative dataset may in fact contain some novel interactions that have not been yet reported. This may suggest that the specificity of PIPE2 may in fact be better than the one observed here. Also it is possible to adjust the filter size and the cutoff to increase the sensitivity at the expense of lowering the specificity as shown in Figure 5.3. For example if we are willing to accept 90% specificity, we can increase the sensitivity to $\approx$ 55% simply by changing the cutoff for the average score after filtering.

5.5.3 All–Against–All Experiment

With the algorithmic improvements described above and the resulting speed improvement, it became feasible to scan the entire *S. cerevisiae* genome for possible interactions. Each of the 6,304 proteins was assessed against all others, resulting in a total of 19,867,056 possible pairs. Every pair was evaluated using PIPE2 with the optimal median filter design and average threshold applied. The experiment was run on a cluster of 38 dual–processor nodes (76 x Intel Xeon 2.0 GHz, 1.5GB RAM) and took

roughly 48 hours to complete. Note that, unlike experimental techniques such as TAP tag, it was possible to evaluate all possible protein pairs, including proteins not amenable to tagging. The resulting interactome is compared with those generated by previous Y2H and TAP tag studies. The predicted PPI's are also evaluated in terms of sub–cellular localization, process and function from GO Slim annotation obtained from the Saccharomyces Genome Database (SGD) [27]. Table 5.3 lists how the GO Slim categories for localization were collapsed.

Group Name	GO Slim compartments
Cytoskeleton	Cell cortex, Cytoskeleton, Microtubule organizing center
Cell wall/bud and site of polarization	Cellular bud, Cell wall, Site of polarized growth
Chromosome	Chromosome
Cytoplasm	Cytoplasm
Membrane	Cytoplasmic membrane–bound vesicle, Endomembrane system, Endoplasmic reticulum, Membrane, Membrane fraction, Plasma membrane
Organelles (besides nucleolus)	Golgi apparatus, Mitochondrion, Peroxisome, Ribosome, Vacuole
Nucleolus	Nucleolus
Nucleus	Nucleus

Table 5.3: GO Slim compartment groups. The groups not shown here did not associate with any yeast proteins and have therefore been omitted from analysis.

5.5.4 Genome–Wide PPI Predictions

We ran all 19,867,056 possible pairs of *S. cerevisiae* proteins through PIPE2 in order to evaluate all possible interactions. This resulted in 29,589 pairs detected as positive interactions. Of these, a slight majority 15,151 (51.2%) have been previously reported, leaving 14,438 as novel interactions that have not been previously reported in any of the databases of interacting proteins (DIP), SGD [21], or BioGRID [61] at the time of the experiment (January 2007). Interestingly, since then, 373 of our 14,438 novel protein interactions (2.6%) have been added to BioGRID. Table 5.4 compares the total number of interactions, average and maximum degree of nodes (interactions

for each protein) and the number of unique proteins participating in interactions according to PIPE2, Gavin *et al.* [36], Krogran *et al.* (S7) [61], Ito *et al.* [51] and Uetz *et al.* [127].

Database	Number of Interactions	Avg. Degree of Nodes	Max. Degree of Nodes	Number of Unique Proteins
PIPE2	29589	11.50	434	5147
Gavin *et al.* [36]	7698	8.90	89	1708
Krogan *et al.* [61]	7123	5.25	141	2708
Ito *et al.* [51]	4007	2.72	239	2944
Uetz *et al.* [127]	862	1.82	21	927

Table 5.4: Comparison of interaction maps between PIPE2 and other high–throughput studies. PIPE2 predicts more interactions using more unique proteins than the other compared methods.

Compared against the TAP tag studies, we observe that proteins in the PIPE2 dataset have slightly more interaction partners on average (11.9) than Gavin *et al.* (8.9) and approximately double the average found in Krogan *et al.* (5.25). It is also important to note the significantly increased number of unique proteins found in the PIPE2 dataset compared to Gavin *et al.* and Krogan *et al.* (approximately 3– and 2–fold, respectively). Similarly, when compared against Y2H studies, the PIPE2 dataset contains almost twice the number of unique proteins compared to Ito *et al.* and over five times more than Uetz *et al.*. These observations may demonstrate one of the strengths of the PIPE2 approach: some PPIs that could not be processed by experimental methods can still be investigated by PIPE2.

5.5.5 Comparing PIPE2 Data to Those Obtained by Genome–Wide Experimental Approaches

It has been previously explained that the overlap between various interaction maps obtained using different methods is very small [4, 23, 34, 127]. A comparison study carried out by Aloy and Russell in 2002 showed a low level of overlap among two–hybrid, affinity purification, mass spectrometry, and bioinformatics methods [4]. Figure 5.4

shows the overlap between PIPE2 data and those of other genome–wide experimental studies.

PIPE2 identifies 96.3% of Ito *et al.* [51] and 91.3% of Uetz *et al.* [127] reported interactions, while Uetz *et al.* cover only 4.32% of Ito *et al.* and 20.1% vice versa (Figure 5.4(a). Figure 5.4(b) presents the overlap between the PIPE2 results and TAP tag studies by Krogan *et al.* [61] and Gavin *et al.* [36]. PIPE2 covers 48.6% of the interactions in Gavin *et al.* and 23.0% PPIs reported by Krogan *et al.* Gavin *et al.* contains 23.9% of Krogan *et al.* and 21.9% vice versa. The diagonal bold numbers in the table show the number of interactions in each database. Exclusion of PIPE2 data highlights the little overlap between the other databases especially between the data obtained by Y2H and TAP tag methods. For example Gavin *et al.* contains only 2.89% of the interactions found in Ito *et al.* and 1.53% vice versa.

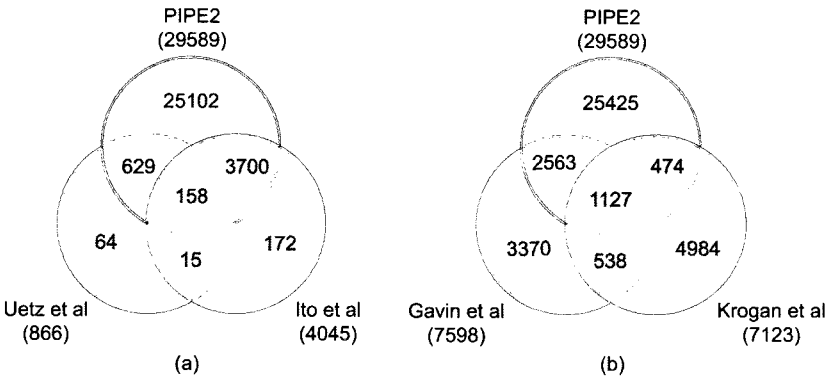

Figure 5.4: Comparing PIPE2 data to those obtained by A) Y2H and B) TAP tag experiments. The overlaps represent the number of interactions which are common between different databases. There seems to be a significant overlap between PIPE data and those of others. This overlap is even more notable for the data gathered using Y2H, which is similar to PIPE2, and designed to study an interaction between two target proteins.

5.5.6 Overlap Between PIPE2 Data and Those of Other Large–Scale Computational Experiments

Recently, other large–scale computational PPI experiments were published such as InSite [131] and Betel *et al.* [15] that attempt to predict PPIs in yeast. InSite bases

its predictions on a set of affinity parameters between pairs of motifs or domains for the query proteins. The published InSite database contains 78,181 protein interactions between 4,450 proteins. The lack of a clear specificity for InSite however, makes the interpretation of this database very difficult. The Betel *et al.* method uses domain-motif interactions based on structure templates of domains of interest. Their database contains 18,458 interactions between 2,311 proteins. As indicated in

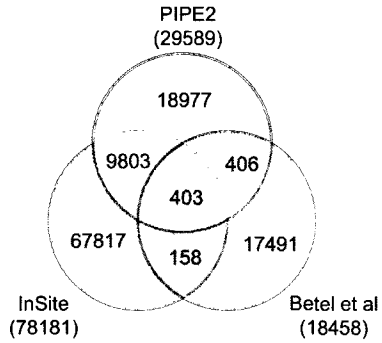

Figure 5.5: Comparing PIPE2 data to those obtained by other large-scale computational experiments. 34.5% of PIPE2 database is found in InSite but only 2.7% of PIPE2 database is shared with Betel *et al.* InSite and Betel *et al.* also share little overlap with 0.7% of Insite found in Betel *et al.*

Figure 5.5, PIPE2 data seems to have a better overlap with InSite but less so with Betel *et al.* (34.5% and 2.7% respectively). This may not be a surprising observation as the method behind InSite that uses affinities between different motifs, has more resemblance to that of PIPE2 that uses re-occurrence of short polypeptide sequences. For the most part Betel *et al.* utilizes predefined domains within structural data with limited availability of detailed binding information. This may explain the small overlap between Betel *et al.* and PIPE2 which does not utilize such predefined information. It should be noted that InSite also has very little overlap with Betel *et al.* (0.7%).

5.5.7 Cellular co–localization of predicted interactors

A summary of the interactions having both proteins localized to the same sub–cellular compartment is shown in Figure 5.6. Localization information in the form of GO

Slim annotation was obtained from SGD [27]. A total of 9,412 pairs (31.8%) were co–localized to the nucleus (38.9%), cytoplasm (25.4%), organelles (except nucleus, 15.19%) and membrane (9.84%). The remaining pairs ($\approx$ 10.7%) were located elsewhere. Therefore, according to PIPE it seems that the majority of the PPIs take place in the nucleus followed by the cytoplasm in a cell. Figure 5.7 also shows a comparison of these numbers with Gavin *et al.* [36], Krogan *et al.* [61], Uetz *et al.* [127] and Ito *et al.* [51] which indicates that the overall pattern for co-localized protein pairs is very similar for all the datasets including PIPE2. Figure 5.8 illustrates the

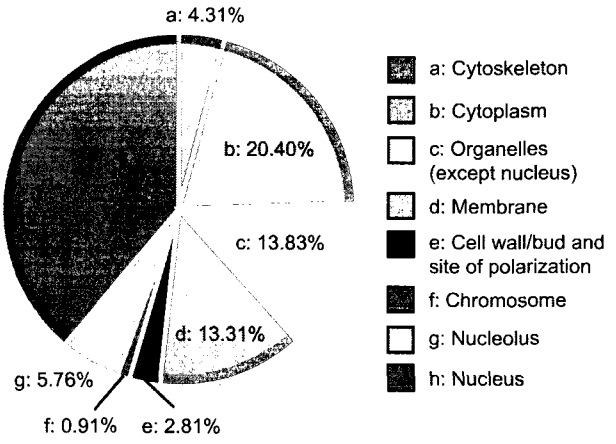

Figure 5.6: Co–localization of PIPE2 predicted interactors.

absolute number of co–localized interactions across PIPE2 predictions in comparison with other large–scale experiments. The total number of co–localized pairs for each dataset is as follows: 9,412 for PIPE2, 3,692 for Gavin *et al.*, 3,283 for Krogan *et al.* (S7), 348 for Uetz *et al.* and 1,435 for Ito *et al.* (see Table 5.5). Furthermore, Figure 5.8 shows for each location the number of novel co-localized PIPE2 interactions in comparison with the number of previously known co-localized interactions (union of other datasets), which is now reported by PIPE2. A large number of novel co-localized interactions are found for the nucleus and cytoplasm. PIPE2 generates more co–localized interactions than the experimental methods in seven out of eight categories. PIPE2 predicts almost 4.5 times more interactions in membrane over that with the second highest count (Ito *et al.*). The percentage of interacting pairs which had different (non-matching) locations in the PIPE2 interaction list is approximately

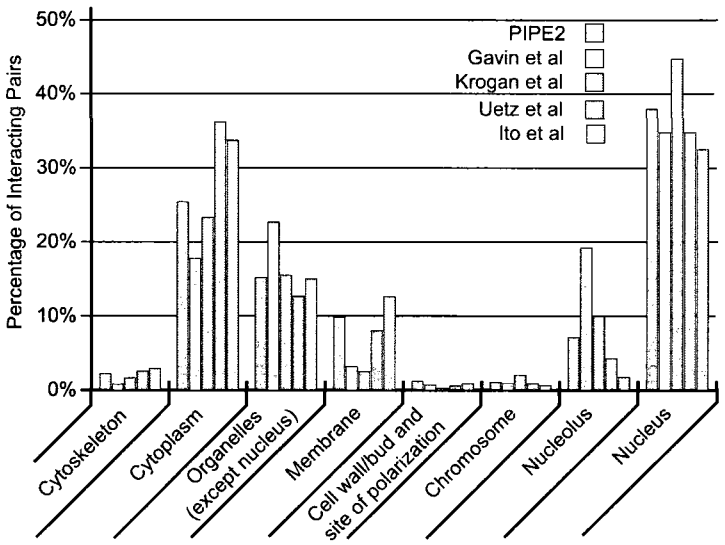

Figure 5.7: Co-localization percentage of predicted interactors for PIPE2 and high throughput experiments. The overall pattern for co-localized protein pairs is very similar for all datasets.

38.4%. This is similar to the other experimental datasets: 33.9% for Gavin *et al.*, 30.17% for Krogan *et al.*, 34.51% for Uetz *et al.* and 36.65% for Ito *et al.* Due to the incomplete and error-prone location data there is no reason to suspect that these interactions are any less valid than the co-localized interactions.

5.6 Investigating the Validity of the Identified PPIs

Interacting proteins generally participate in functionally related processes [58, 135]. Consequently, sharing functional properties may provide further validations for the predicted interactors. To investigate the validity of observed interactions, we randomly selected three sets of one hundred (3×100) protein pairs from the 14,438 novel PIPE2 interactions. We then investigated primary literature to manually determine the common functional information for each pair. The results of this analysis are shown in Table 5.6. It was observed that 20, 22 and 17 interacting pairs, in the three selected sets of novel protein pairs, respectively, also had a previously reported functional relationship. For example, the novel interaction detected for YDL213C and

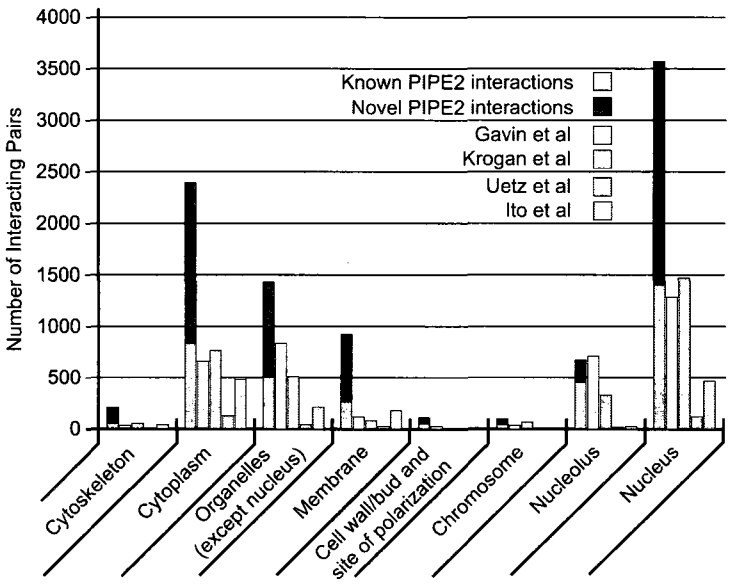

Figure 5.8: Co-localization of predicted interactors for PIPE2 and high throughput experiments. The location of interacting pairs does not seem to affect the ability of PIPE2 to predict an interaction. Besides nucleolus, PIPE2 predicted more interactions in all other cellular locations. PIPE2 detected almost 4.5 times more interactions in membrane than the second highest count (Ito *et al.*). PIPE2 data is divided to two categories of novel, and those that overlap with the union of others.

YKR092C was supported by their previously reported functional roles in ribosome assembly [21]. Hence, 59 of the possible set of 300 novel interactions detected by PIPE2, or 20%, can also be supported by a functional relationship. Similarly, a second potential line of validation for an interaction might be that the interacting proteins often have common interactors (common third protein interaction) [19, 94]. Our manual survey of interaction databases from primary literature indicated that 49, 45 and 46 protein pairs among the three selected sets of novel interactions, respectively, had previously reported common interactions with at least one other protein (Table 5.6). Hence, 140 of the possible set of 300 novel interactions detected by PIPE, or 47%, can also be supported by a previously reported common interaction. For example it was observed that the two novel interacting proteins YKL068W and YML007W both had previously reported interactions with three common proteins YER110C,

Location	PIPE2	Gavin *et al.*	Krogan *et al.*	Uetz *et al.*	Ito *et al.*
Cytoskeleton	211	29	54	9	42
Cytoplasm	2392	655	765	126	484
Organelles (besides nucleus)	1430	837	508	44	215
Membrane	926	118	84	28	181
Cell wall/bud and site of polarization	110	26	9	2	13
Chromosome	100	34	75	3	9
Nucleolus	672	710	327	15	25
Nucleus	3571	1283	1469	121	466
Total	9412	3692	3283	348	1435

Table 5.5: Comparison of interaction maps between PIPE2 and other high–throughput studies.

YGR218W, and YMR047C [2, 51, 137]. Interestingly these three common interactors are all involved in a similar process of nuclear protein transport. Altogether, 39 of the 300 novel pairs, or 13%, were supported by both a functional relationship and the presence of a third common interacting partner (Table 5.6). Similarly 199 of the 300 interactions, or 66%, were supported by at least one of the investigated additional lines of evidence. We then include GO Slim annotation [27] from the SGD

Set Number	Evidence based on function	Evidence based of 3^{rd} interaction	Both lines of evidence	Either lines of evidence
1	20	49	17	69
2	22	45	13	67
3	17	46	9	63
Total	59	140	39	199

Table 5.6: Analysis of PIPE2 data using functional relationship and the presence of a common interactor, using data from primary literature.

database to investigate the entire set of 14,438 novel PPIs detected by PIPE2 for the presence of a relationship between the interacting partners which may support the

validity of an interaction (Figure 5.9). The investigated information included sub-cellular localization (compartment), cellular process, molecular function and common third party protein interaction. Each of these common features is represented by a different circle in Figure 5.9 and the overlaps indicate the number of pairs that share additional features. As indicated in Figure 5.9, 8,712 novel interactions (total number of interactions shown in Figure 5.9), or 60% (calculated by 100 X [number of novel interactions with at least one common feature]/[total novel interactions] or 100 X 8,712/14,438), possess at least one additional common feature for the novel interacting proteins. Similarly, 3,319 (overlaps between two or more circles), 885 (three or more circles), and 148 (four circles) protein pairs showed at least 2, 3 and 4 common features, respectively (Figure 5.9). These categories of the novel PPIs may also be used by researchers to prioritize their confidence in the predicted interactions.

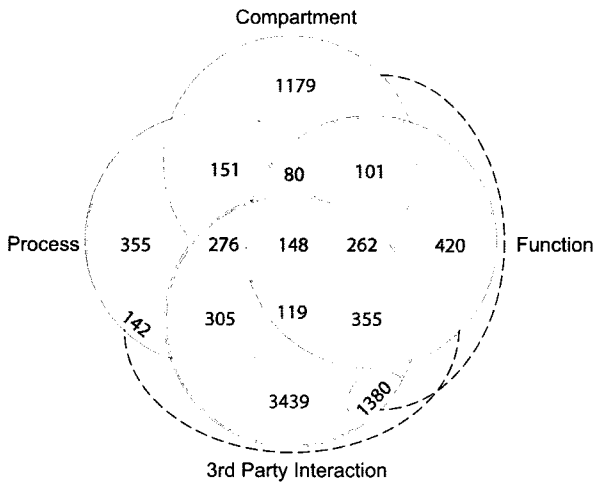

Figure 5.9: Analysis of PIPE2 novel interactions by compartment, function, process and third common protein interaction. A different circle represents each feature. Protein pairs represented here (8,721) indicate those that share at least one common feature (of 14,438 total novel pairs). Overlaps represent additional common features. Dashed lines connect overlapping areas for compartment-third party interaction and process-function.

5.7 PIPE2 Data Can Reveal Novel Protein Complexes

Protein complexes are formed from the interaction of two or more functionally related proteins to carry out a specific cellular function. More than 500 protein complexes have been previously reported in yeast [98]. It is estimated that this number might in fact be closer to 800 [40]. Consequently, there may remain many complexes which are yet to be identified. Here we have identified over 14,000 novel interactions. Therefore it might be expected that this information can be used to determine new members of previously determined complexes or to discover novel protein complexes. In particular, membrane proteins often provide a challenge for the experimental PPI identification methods. PIPE2 novel predictions revealed that four characterized or putative membrane proteins belonging to the family of DUP240 proteins, YGL051W, YAR027W, YAR028W and YCR007C interact with each other and with four other proteins YAR033W, YOR307C, YLR065C and YKL174, and form a 4–member core for a complex of 8 interacting proteins. DUP240 proteins form a family of trans–

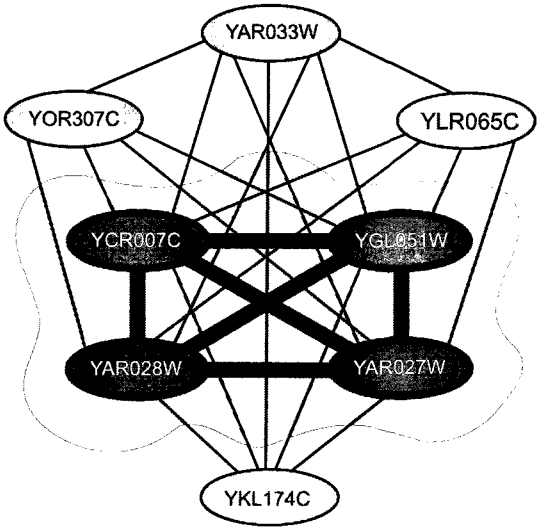

Figure 5.10: A novel yeast complex revealed from PIPE2 data. YAR027W, YGL051W, YCR007C and YAR028W interact with each other as well as all other proteins and may represent the core of the complex. Besides YLR065C, which is uncharacterized, all other proteins are thought to be involved in vesicle formation or function. Consequently, these interactions may also suggest a vesicle associated role for YLR065C.

membrane proteins, which are believed to be involved in vesicle formation [96, 107]. YAR033W is another member of the DUP240 family. YOR307C and YKL174C are two vesicle–associated proteins, and YLR065C is an uncharacterized open reading frame of unknown function. Such common functional properties of these proteins may provide further support for the predicted interactions. In addition, it is well–documented that functionally related proteins often interact with each other [19, 58]. Consequently, based on the novel interactions that PIPE2 predicted for YLR065C, it can be hypothesized that YLR065C may have a putative role in vesicle formation or function. In agreement with this hypothesis, it has been shown that YLR065C in combination with either YGL020C or YER083C results in a synthetic lethal genetic interaction, both of which are reported to be involved in retrograde vesicle–mediated transport from Golgi to endoplasmic reticulum [87, 108].

5.8 Novel Information Extracted From PIPE2 Data

Non–homologous end–joining (NHEJ) is a DNA repair mechanism by which the two ends of a double–stranded DNA break (DSB) rejoin in the absence of a significant homologous template. A number of different proteins and processes have been shown to affect the efficiency of NHEJ in *S. cerevisiae* and other eukaryotes. This number is continuously growing [54, 60, 78], suggesting that there remain other undiscovered proteins capable of affecting the efficiency of this DNA repair mechanism. It is well established that functionally related proteins often interact with each other. Therefore, to further investigate the biological relevance of PIPE2 data, we studied PIPE2's novel PPIs to discover potential gene candidates that may be involved in NHEJ. We observed that YDL012C and YOL012C form novel interactions with YMR106C (yku80), a key factor in NHEJ [24] and YLR442C (Sir3), also known to affect the efficiency of NHEJ [123]. YDL012C is an uncharacterized open reading frame and YOL012C is a histone variant involved in regulation of transcription and chromatin silencing [66]. Neither of these proteins has been directly linked to NHEJ. Based on the observed interactions we hypothesized that these two proteins might be involved in NHEJ. To examine a possible role for YDL012C and YOL012C in

NHEJ, we subjected their gene deletion yeast strains to a plasmid repair assay analysis [52, 78]. In this approach, target and control yeast strains are transformed in parallel with the same amount of intact and linearized plasmids. The region around the site of linearization, which is generated by restriction digestion, has no homology to the yeast chromosome. Consequently, those strains that have compromised NHEJ show deficiency in circularizing the plasmid and thus form fewer colonies on a selective media. It was observed that in the absence of YDL012C and YOL012C, yeast cells had reduced efficiency in repairing linearized plasmid (Figure 5.11). These observa-

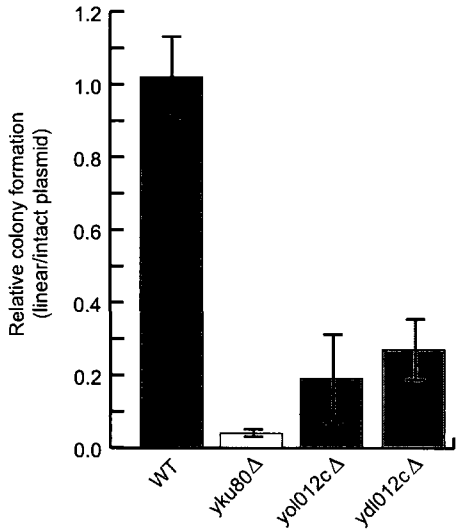

Figure 5.11: Plasmid repair efficiencies of yeast deletion mutants. The ratio of the number of colonies formed after transformation with linearized plasmid to that formed with the intact plasmid is used to represent the efficiency of NHEJ. This ratio for yku80Δ, yol012cΔ and ydl012cΔ is 0.04±0.01, 0.19±0.12 and 0.27±0.08, respectively. Each experiment is repeated at least four times. Yku80 is a key factor in NHEJ and its deletion strain (yku80Δ) is used as a positive control. WT is the wild-type strain.

tions suggest that both YDL012C and YOL012C affect the efficiency of NHEJ. These data further indicate that novel biologically meaningful information can be extracted from PPI data gathered by PIPE2. It should also be noted that further investigation is required to elucidate the mechanism by which these two proteins affect NHEJ.

5.9 Summary

The 16,150x fold performance increases to the PIPE program made possible the first all–to–all sequence comparison based on re–occurring motifs leading to over 14,000 new interactions. PIPE2 can on average predict the interactions for two protein pairs per second, which is substantially faster than the previous one pair per two hours. Running an all–to–all experiment on *S. cerevisiae* ($\approx 6,300$ proteins, $\approx 20M$ pairs) with the original PIPE was expected to take over 60 years even when using a cluster of 76 processors (over 4,500 years using one processor). With the improved speed of PIPE2 we are now able to complete this task in approximately 2 days of computation time. The reduced computational requirements also allowed us to refine our method by running thousands of pairs of known positives and negatives instead of being confined to our initial sets of 100 positives and 100 negatives. It also enabled us to revise our threshold function for determining whether or not a pair interacts thereby increasing the specificity to 99.9%. This is critical when running a large number of pairs since a large number of false positives will be generated even if the specificity is relatively high. When evaluating only 100 pairs, the 89% specificity of the original PIPE is expected to generate 11 false positives, but for 2M pairs the original PIPE would have reported approx. 2,200,000 false positives, which is unacceptable for large–scale investigations. PIPE2 solves this problem.

Previous genome–wide analyses of PPIs have predominantly relied on Y2H and TAP tag methodologies. These techniques are both time and labor intensive and they both have high rates of false positive and false negatives results associated with them ($\approx$ 45% false positive rate for Y2H and 15–50% false positive rate for TAP [29]). These two methods have inherent advantages and disadvantages. For example, TAP tag obtains interaction information from a more natural environment since the physiological conditions are more realistic than those created by Y2H. However, like all affinity purification methods, TAP tag also suffers from co–purification of contaminants. Additionally, it is often difficult to differentiate between direct and indirect (i.e. via a third partner) interactions using TAP tag. Another common limitation for these techniques is that they cannot be applied to all proteins without discrimination. In TAP tag, the double tag fusion to the target protein may interfere with the

formation of some complexes or cause a mutant phenotype [35, 133]. In Y2H, not all proteins can be safely over–expressed and not all proteins can find their way into the nucleus which is required for the successful detection via Y2H [58]. Such limitations resulted in small overlaps between the PPI data collected using different approaches and even little reproducibility using the same method in different experiments [35]. This lack of overlap suggests the presence of more undiscovered PPIs.

Here, using re–occurring short polypeptide sequences between known interacting protein pairs, we systematically investigated all possible interaction pairs (approximately 20M) between all yeast proteins. In this way we have identified 29,589 PPIs, approximately 15,000 of which were previously reported interactions. The overall overlap between PIPE2 data and the data from other large–scale investigations are very similar to the overlap observed between previous experimental procedures (See Figure 5.5). This is somewhat expected considering the differences in the methodologies applied in these investigations. The observed average node degree of 11.5 identified by PIPE2 does not differ greatly from the average of 8.91 observed by Gavin *et al.* [36]. It is, however, significantly higher than those observed in other genome–wide investigations. This is not a surprising observation given the higher overall number of interactions observed with PIPE2.

PIPE2 identified approximately 14,000 novel interactions. This may stem from the ability of PIPE2 to investigate all proteins without discrimination. This is a major advantage of PIPE2 over TAP tag and Y2H methods where not all proteins can be subjected to analysis (see above). An example of this is seen in Figure 5.8, where a significant number of membrane proteins have been identified in the PIPE2–generated interactome. Because of their inherent properties, applying TAP tag and Y2H analysis to membrane proteins has proven to be challenging. PIPE2 analysis also had a high level of success in identifying PPIs in the nucleus. This might be explained by the presence of a high number of essential proteins in the nucleus, which may not be readily manipulated by Y2H or TAP tagging. An area where PIPE analysis had a relatively low level of success was for nucleolus proteins; see Figure 5.8. It appears that TAP tag experiments by Gavin *et al.* [36] had a significantly higher relative success in identifying these interactions. The nucleolus is the site of ribosomal

RNA synthesis and biogenesis. One possible explanation therefore may be that the relatively high number of protein complexes at work in this region (both protein–protein and protein–RNA–protein) may result in an inflated number of interactions detected by TAP tag. This is mainly due to the inability of TAP tag to readily differentiate between direct and indirect (via a third partner) PPIs.

A significant limitation of PIPE2 is that it relies exclusively on a library of pre–existing experimentally–derived interaction data for identifying re–occurring short polypeptide sequences. Consequently, in the absence of sufficient data for an interacting short polypeptide sequence pair, PIPE2 will be ineffective. Due to the window length of 20AA, it is also possible that short interaction sites will be missed or overlooked. PIPE2 will also be less effective for motifs that span discontinuous primary sequence, as it does not account for gaps within the short polypeptide sequences. It is expected that the use of more refined algorithms that permit such gaps, along with an increasing number of available libraries of PPIs may increase the accuracy of PIPE2.

Increasing availability of three–dimensional protein structures may also provide an improved starting dataset for PIPE2 analysis, which may result in a further increase in the accuracy of this tool. Another possible future direction is to reduce the rate of false positives by incorporating vigorous filters that consider other information about the target protein pairs, including sub–cellular localization or functional annotation.

Chapter 6

PIPE3: Prediction in Other Organisms

6.1 Introduction

Traditional high–throughput attempts, which are both time and resource demanding, have been utilized to uncover the global PPI networks of only a few model organisms. The PPI networks in other organisms including human has remained mainly unexplored. Here we show that the presence of re-occurring short polypeptide sequences between interacting protein partners seem to be conserved over the course of evolution, providing an opportunity to predict PPIs from primary sequences of proteins alone, in organisms such as human. As a proof of principle, we predict and analyze hundreds of novel human PPIs, experimentally confirm predictions, identify PPIs between human proteins and viral pathogen proteins, determine the potential interaction sites of proteins, and predict an all–to–all PPI interaction map for *S. pombe* ($\approx 9,000$ PPIs). The results provide an effective alternative to experimental methods for high throughput examination of PPIs. Further implementation of this approach can lead to the prediction of thousands of more novel and testable PPIs in different organisms and will help future studies on individual proteins and systems biology.

We've already seen that the PIPE algorithm can be useful to predict PPIs in *S. cerevisiea*. PIPE2 improved the running time and accuracy of the PIPE algorithm. In this chapter we will introduce the third iteration of the PIPE algorithm, aptly named PIPE3. With this new version we investigate the possibility to extend the PIPE method to other organisms such as *S. pombe*, *E. coli* and ultimately *H. sapiens*. An online PIPE3 portal has been made available at: http://cgmlab.carleton.ca/PIPE3/.

6.2 Predictions In Other Organisms

As previously discussed, PIPE has been trained on *S. cerevisiae* in order to predict interactions in that organism. However the algorithm itself is not dependent on that organism. To be able to predict protein–protein interactions in another organism, all of PIPE's parameters would need to be re–assessed and the database replaced with one containing known interactions for each particular organism. The expected score of two fragments would need to be tested in order to set the S_{pam} parameter. The scoring matrix used might have to be changed in order to better match the goal organism. The score random sequences would need to be determined to set the score threshold. Finally, instead of using the median filter and average score cutoff used for *S. cerevisiae* another type or size of filter might be needed.

If the algorithm could be made to successfully predict interactions in any organism for which the genome has been sequenced it would be an invaluable tool to predict the complete interaction map for that organism. For new organisms it would offer biologists a guide to follow in order to experimentally detect the interactions using traditional methods. This is the target of PIPE3: predictions in organisms other than *S. cerevisiae*. We first had to pick which organisms would make good candidates for testing. The following organisms of interest were targeted:

- *Schizosaccharomyces pombe*: also called "fission yeast", it is somewhat similar to *S. cerevisiae* so it becomes a logical next step in testing PIPE for other organisms. It is estimated to contain 4,970 genes, less than *S. cerevisiae*. It also has relatively few known interactions: $\approx$ 3,000.

- *Escherichia coli*: a bacteria containing $\approx$4,300 genes. *E. coli* has been studied extensively and is another model organism for lab experiments. Some strains are harmless while others can cause severe food poisoning in humans. One large–scale experiments for the *E. coli* K–12 strain has identified 2,667 PPIs [7].

- *Caenorhabditis elegans*: a small free-living nematode (roundworm) which has been used extensively for interaction studies. Their genome encode for roughly

19,000–20,000 genes which makes it more complex than the *Drosophila*. However the largest protein map to date only includes $\approx 5,500$ interactions [68].

- *Homo Sapiens*: the ultimate goal of the PIPE project is to predict accurate interactions involving human proteins. The human genome is believed to code for 20,000–40,000 proteins (there is still some debate on the actual number) and is estimated to have between 154,000 and 369,000 interactions [46]. The Online Predicted Human Interaction Database currently contains 47,221 interactions involving 10,579 unique proteins [18]. Even if we assumed every predicted interactions is accurate that would still only represent 1/3 of the complete network map in the best case ($\approx$ 13% in the worst case). The large number of proteins in human provides a computational challenge for any proteome-wide computational method. All possible pairs using 25,000 proteins would yield over 300 million pairs. Using the same cluster used to perform the genome-wide scan of the yeast genome and assuming the average protein length was the same it would take roughly 15x the runtime ($\approx$ 30 days). Since human proteins are longer on average than yeast proteins that runtime could be more than doubled.

- *PPI in other organisms*: the list of organisms above represents a small list of possible organisms possible for PIPE. Depending on data availability many other organisms could be included for future testing such as: *Mus Musculus*, *Helicobacter pylori*, *Drosophila melanogaster*. By combining interaction databases from many different organisms we will see how it is possible to create a database which could be used with any organism even in the absence of interaction data for that particular organism.

6.3 PIPE3 Parameter Tuning for Other Organisms

Due to the different amino acid composition of each organism, some parameters have to be tuned before making predictions in other organisms. One such parameter is the score cutoff for window fragment comparisons. Since the window comparison is based on a substitution matrix, a different amino acid composition will cause the expected window score to shift slightly. In Figure 6.1(a) we see the histogram of score

distribution for random fragments of length 20 (window length) changes somewhat for different organisms. In Figure 6.1(b) we plot the probability of scores when comparing two fragments of length 20 and we pick our threshold to be significantly above random (probability of 10^{-6} of getting this score by comparing random fragments). We have previously determined [94] this cutoff to be 35 in *S. cerevisiae* and that remains a valid cutoff for most of the other organisms tested except one: *H. sapiens*. For *H. sapiens* the expected random score when comparing fragments increased, so a new cutoff of 40 was used exclusively for that organism.

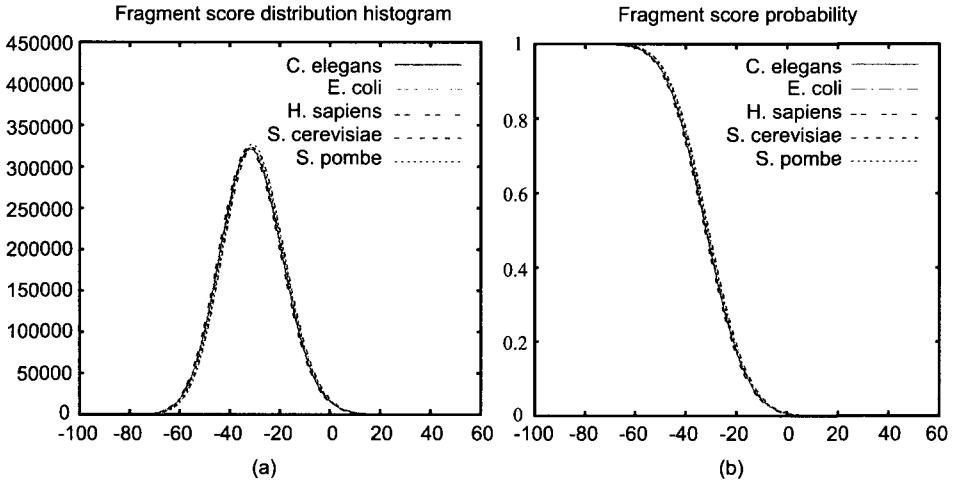

Figure 6.1: (a) Histograms of fragment scores for all organisms tested. (b) Probabilities of fragment scores for all organisms tested. A cutoff equivalent to 10^{-6} probability was chosen for each organism.

Changing the fragment window length (20 AA) and substitution matrix (PAM120) did not offer any significant improvement. For any new organism the sequence of operations to tune and set the PIPE parameters are as follows:

1. Gather all the protein sequences for the organism (usually from an online source). We can limit the proteins to those involved in known interactions, however the entire genome is preferred.

2. Build a set of known interactions for the organism. This can be from one experiment, or from a repository such as BioGrid [119]. It can also be a gold-standard set or the union of all the databases online for example. Only physical

interactions should be used since this is what the PIPE algorithm is meant to predict.

3. Calculate the amino acid distribution for the organism. This is then used to evaluate the expected score for comparing two random fragments of length 20 using that distribution (see Figures 6.1 and 4.2 for examples). We choose the S_{pam} score which gives us a 10^{-6} probability of getting a random match.

4. Once we have the correct settings we can run the database pre-computation as explained previously. The pre-computation checks every window of length 20 against the entire database and records the matches it finds.

5. With the pre-computation done, we must evaluate the sensitivity and specificity of the predictions for this organism. This is done by LOOCV using the known interaction set collected earlier. The negative set is normally a set of random pairs ([13]). Plotting the ROC curve of the LOOCV experiment will indicate what cutoff to use in order to achieve a given specificity or sensitivity. For high–throughput experiments, a high specificity is suggested in order to avoid a large number of expected false–positives.

6. After completing all these steps we are ready to start predicting new PPIs. For small genomes this can be done by an all–to–all experiment but for large genomes it is often preferable to only run pairs of interest due to the runtime involved.

6.4 Results of Predicting PPIs in Other Organisms

To study the evolutionary conservation of interaction codes, we first investigated whether PPIs in other organisms may also be predicted from their primary sequences. We used the available information about the PPI pairs in five model organism, *E. coli, S. pombe, C. elegans, S. cerevisiae* and *H. sapiens*, to predict interactions in each respective organism (ex: known *S. pombe* interactions to predict novel *S. pombe* pairs). Table 6.1 presents the number of protein sequences, the number of physical interactions and the interaction database used for the organisms tested. In an effort to

have the most accurate information, our interaction list for *S. cerevisiae* was updated to include the most recent interaction list. The sensitivity and specificity for *S. cerevisiae* will therefore be slightly different than those presented in Chapter 5.

Organism	Number of Proteins	Number of Interactions	Interaction Database
S. cerevisiae	6,716	43,591	BioGRID [119]
C. elegans	23,684	6,607	BioGRID [119]
E. coli	13,179	16,235	EciD [6]
H. sapiens	22,513	41,678	HPRD & BioGRID [55, 119]
S. pombe	5,024	2,951	BioGRID [119]

Table 6.1: Number of protein sequences, physical interactions and the interaction database for the organisms tested. Only physical interactions are extracted from the respective database for each organism.

Apparent from the areas under the curve (AUC, Figure 6.2), our plotting of the fraction of true positives, over false positives (ROC curves) suggests that it is possible to successfully detect PPIs, from the interaction codes, in all five organisms that we investigated. At the higher specificities (99.9 - 99.95%), which is what is needed for genome-wide analysis, it appears that humans have the highest accuracy followed by *S. pombe, C. elegans*, and *S. cerevisiae*. For example, at a specificity of 99.95%, 23.8% of human PPIs can be predicted. At the lower specificities however, for example when predicting interactions between a few target proteins, the accuracy for *S. pombe* surpasses that in humans. The accuracy for *E. coli* at high specificity is very low suggesting that *E. coli* predictions are only possible at lower specificities.

Table 6.2 presents the achieved sensitivity using relatively high specificities (90.0%, 95.0%, 99.0% and 99.95%) for the different organisms tested and different versions of PIPE. As with Figure 6.2, we can see that at very high specificities (99.95%) *H. sapiens* has the best sensitivity (23.8%). At the lowest specificity shown (90.0%), *S. pombe* performs better than all the other organisms (77.5%) which is also visible in 6.2.

Next, we investigated if PPIs in a target organism can be predicted from cross-species PPIs (Figure 6.3). Specifically, can we predict interactions in one organism

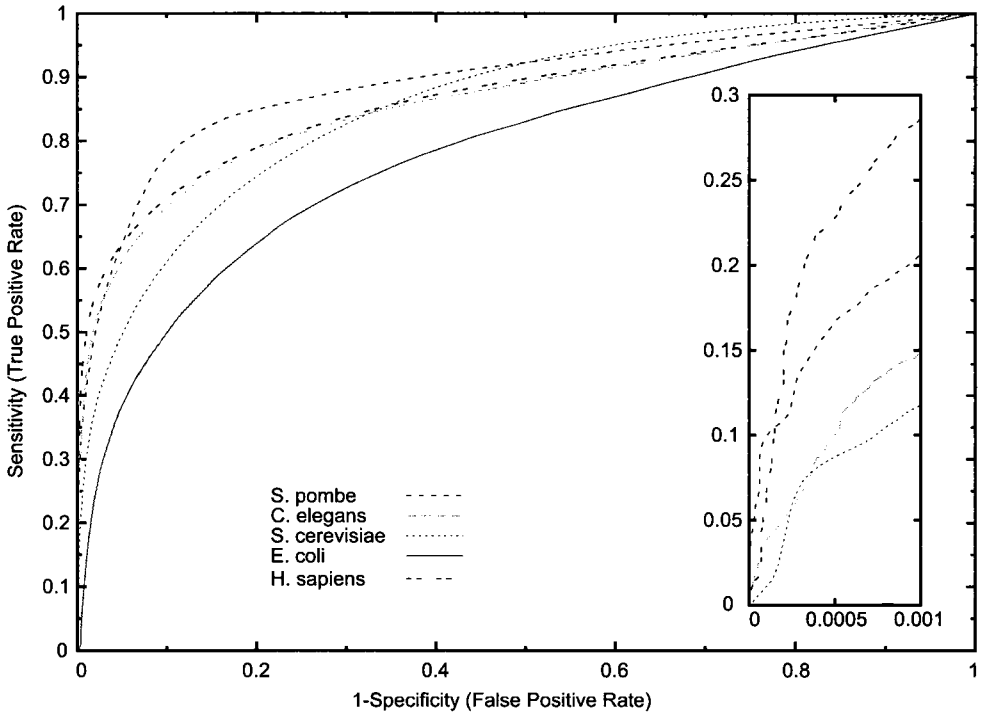

Figure 6.2: ROC (Receiver Operating Characteristic) curves for *C. elegans, E. coli, H. sapiens, S. cerevisiae* and *S. pombe*. The curve presents the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity). *Insert:* at very high specificity (99.95%) *H. sapiens* has better sensitivity than all other organisms tested.

using known PPIs from a different organism; i.e. predict novel *S. cerevisiae* interactions using *H. sapiens* PPIs and vice versa. We can see in Figure 6.3 that while using known *S. cerevisiae* PPIs to predict *H. sapiens* interactions (SC–HS) is lower than using known *H. sapiens* PPIs to predict *H. sapiens* interactions (HS–HS), we can still gather some meaningful interactions especially at lower specificities (46-48% sensitivity at 80% specificity for both organisms). The implication of this experiment is that we are able to use known PPIs in one organism to predict novel interactions in others with respectable sensitivity/specificity.

Next, we ask the question: "if PIPE can predict novel interactions in one organism using known PPIs in another, how would it perform if the predictions are based on PPIs from multiple organisms?". Therefore we extended our previous investigation to determine if PPIs in a target organism can be predicted from cross-species PPIs using

Organism	90.0% Sp.	95.0% Sp.	99.0% Sp.	99.95% Sp.
S. cerevisiae (PIPE)*	≈60.0%	≈45.0%	N/A	N/A
S. cerevisiae (PIPE2)	53.1%	40.0%	22.0%	17.3%
S. cerevisiae (PIPE3)	61.1%	49.6%	30.3%	8.8%
C. elegans (PIPE3)	69.7%	61.2%	43.6%	10.0%
E. coli (PIPE3)	49.9%	38.5%	14.5%	N/A
H. sapiens (PIPE3)	72.7%	65.0%	51.1%	23.8%
S. pombe (PIPE3)	77.5%	63.9%	41.3%	6.4%

Table 6.2: Comparing sensitivities achieved at given specificities for all tested organisms and for different version of PIPE. The ROC curves in Figure 6.2 presents how each organism reacts to increasing or lowering the cutoff, but this table presents some points of interest. Here we see the sensitivities achieved by every organism tested at specific specificities: 90.0%, 95.0%, 99.0% and 99.95%. *Since the original PIPE method was only tested with 100 positives and 100 negative pairs (due to runtime constraints), the sensitivity given for 90.0% and 95.0% are approximations and higher specificities (99.0% and 99.95%) could not be tested so are not available (N/A). No possible cutoff resulted in 99.95% specificity for *E. coli* so this result is not available (N/A).

the union of multiple species (Figure 6.4). In this experiment, all known interactions from our select organisms (*C. elegans, E. coli, H. sapiens, C. cerevisiae* and *S. pombe*) are used except the interaction from the organism used for testing. For example to test our prediction in *H. sapiens* we would use all known interactions for the organisms above except interactions for *H. sapiens*. This simulates predictions in a new organism or for one which has few known interactions. Again our ROC curve suggests that the number of true positives over false positives supports the ability to use interaction pairs from different species to predict interactions, in both high and low specificities, in an independent species, for all eukaryotes tested. The efficiencies of these predictions however, are lower than in Figure 6.2 but similar to Figure 6.3 for *H. sapiens* and *S. cerevisiae*. In fact for *H. sapiens* we see a slight sensitivity improvement at the 80% specificity mark (54–55% sensitivity as opposed to 48–49%). Again *E. coli* PPI prediction seems to be an exception, suggesting a difference between protein codes in *E. coli*, a prokaryote, and other organisms examined.

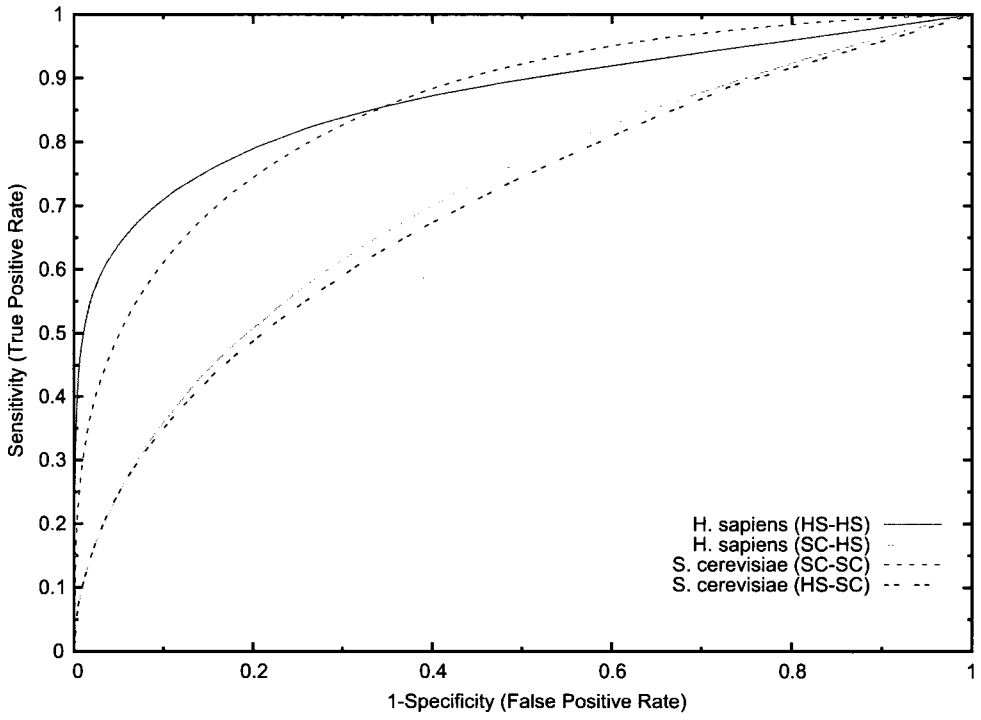

Figure 6.3: ROC curves for cross-species prediction. Here we see the results of using known *H. sapiens* interactions to predict *S. cerevisiae* interactions (HS-SC) and known *S. cerevisiae* interactions to predict *H. sapiens* interactions (SC-HS) compared to same-species prediction (HS-HS and SC-SC).

The runtime for our predictions under our current specification, makes predicting the human global protein interaction map out of reach at this point (≈6.3 million CPU hours or ≈3 years on a 256–cpu cluster). We therefore targeted *S. pombe*, which has 5,024 proteins and has the second best ROC curve in our analysis. In doing so, we detected a total of 9,009 possible interactions, 6,058 of which are novel/false positives. Since currently there are 2,951 known interaction pairs in *S. pombe*, our predictions have increased our knowledge of PPIs in *S. pombe* by over three–fold. To examine the quality of the predicted interaction pairs we classified them according to their molecular function, biological process and location inside the cells (Figure 6.5). The center number in the overlapping circles lists the percentage of pairs that are co–localized AND have the same function AND are involved in the same process.

We can see in Figure 6.5 that PPI pairs predicted by PIPE3 have similar functions,

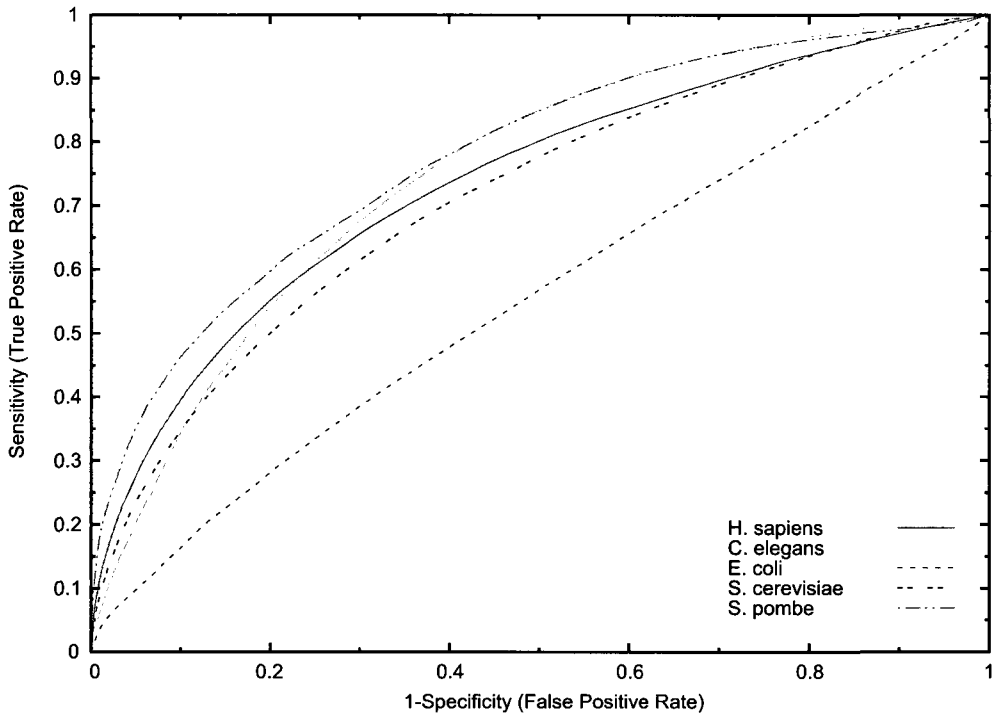

Figure 6.4: ROC curve measuring the prediction accuracy of predicting interactions using the union of known interactions in several other organisms except known interactions for the test organism.

cellular component and process with similar percentages than previously detected interactions. Compared to random, PIPE predictions have almost double the number of co–localized pairs. Third party interactions (where both partners interact with another common protein) were also verified to further assess the quality of the interactions. For random pairs, 0.032% had third party interactions compared to 36.2% for PIPE3 predictions and 76.5% for previously detected interactions. Here again we see a huge difference between the random numbers and PIPE's predictions. The previously detected interactions might have an unfair advantage in these tests: proteins known to interact are then studied and assigned GO terms while PIPE3's predictions can involve proteins for which no information has been found yet.

PPI data can also be used to determine protein complexes. Using our predicted novel PPIs we identified two novel protein complexes in *S. pombe*. One of them is

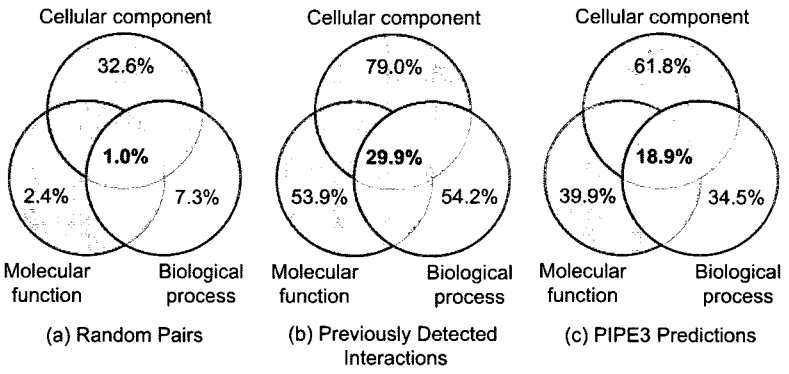

Figure 6.5: Percentages of pairs in which both partners share the same GO Slim annotation in *S. pombe*. (a) Results for random *S. pombe* pairs (100,000 pairs). (b) Results for previously known interactions (BioGRID [119], 2,951 interactions (c) Results for PIPE3 predicted pairs (9,009 predicted interactions).

a five member complex with *SPBC1289.13c*, a putative galactosyltransferase as a core protein that interacts with four other proteins, some of which interact with each other. Two of these proteins, *SPAC22E12.06c* and *SPCC736.04c* are known to have alpha-1,2-galactosyltransferase activities involved in N-linked, and both O-linked and N-linked oligosaccharide modification of proteins, respectively ([138]), and have previously reported to interact with each other. All five proteins are membrane proteins that are associated with Golgi apparatus. Therefore it is likely that this complex has a role in galactosylation of glycoproteins. The second complex consists of 8 members, with 5 proteins forming the core, and 3 additional proteins that interact with the core proteins but not with each other. They are all thought to be membrane proteins. Five of these proteins (*SPBC1703.10, SPAC4C5.02c, SPAC6F6.15, SPAC9E9.07C* and *SPAC18G6.03*) have been linked to protein transport and vesicular trafficking. This suggests a role for the complex in this process.

To evaluate the effectiveness of the PIPE method to predict PPIs in human we investigated the interactions for 29 proteins, with established roles in the efficiency of double stranded (ds) DNA break repair, against all human proteins, representing a total of more than 650,000 possible protein pairs (29 x 22,500). dsDNA breaks represent a severe case of DNA damage which if remained unrepaired, can lead to cancer development. With a specificity of 99.9% we detected a total of 1,657 interactions,

511 of which were previously reported, and 1,146 of which represented novel/false-positive interactions, therefore increasing our knowledge of potential PPIs for dsDNA repair break proteins by more than three–fold.

To investigate the relevance and validity of the identified novel interactions, we further analyzed them based on molecular functions, cellular processes, and co-localization of the interacting partners (Figure 6.6) along with the presence of a common third protein partner interaction. The center number in the overlapping circles lists the percentage of pairs that are co–localized AND have the same function AND are involved in the same process.

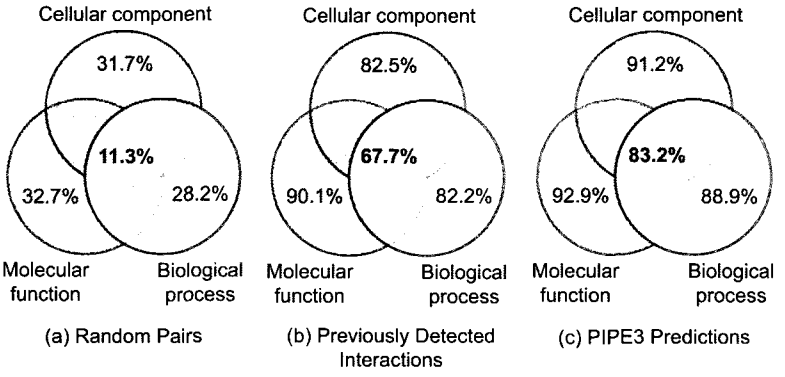

Figure 6.6: Percentages of pairs in which both partners share the same GO Slim annotation in *H. sapiens*. (a) Results for random *H. sapiens* pairs (2,500 pairs). (b) Results for previously known interactions (Biogrid [119], 41,678 interactions (c) Results for PIPE3 predicted pairs (1,056 predicted interactions).

Just as in *S. pombe*, PIPE's predicted pairs share GO terms more frequently than the random set. In this case however the pairs predicted by PIPE3 are actually better than previously detected interactions in each of the three categories. Third party interactions (where both partners interact with another common protein) were also verified to further assess the quality of the interactions. For random pairs, 0.3% had third party interactions compared to 45.3% for PIPE3 predictions and 59.2% for previously detected interactions. We see the same trend as with *S. pombe*: PIPE3 predictions behave close to previously detected interactions than to random pairs.

To further assess the effectiveness and reliability of our approach, we applied our method to predict PPIs in a set of gold standard. Yeast gold standards are commonly

used in yeast PPI investigations and represent interactions for which a significant body of independent evidence is available. We assumed that the interacting pairs within a gold standard that are evolutionary conserved from yeast to human may represent true interactions in humans, and in the absence of a customary human alternative, they may represent a gold standard for evaluation of PPI predictions. Of 701 gold PPIs for which human homologs are found 252 and 347 interactions were predicted at the specificity of 99.9%, and 95%, respectively, resulting in a corresponding sensitivity of 35% and 49.5%. This is a higher sensitivity than we observed above and may be explained by the increased density of true interactions in a gold standard.

The presence of evolutionary conserved interaction regions also implies that predicting interactions between proteins from different species, for example between a host and a pathogen protein, may also be possible. We therefore examined the interactions for proteins of human pathogens *Hepatitis C (HCV)*, *Influenza A*, *HIV-1* and *HIV-2*, with all human proteins. For example, we found novel interactions for HCV non-structural 5A (NS5A) protein with several human proteins previously known to be involved in different non-HCV host-pathogen interactions. These novel interactors include human IPO5 known to mediate nuclear import of *HIV-1 Rev* protein [8], and the proto-oncogene tyrosine-protein kinase SRC, which is involved in host-hepatitis E virus interaction [57]. The predicted novel interactions suggest that these proteins may also be involved in HCV infection. Similarly, we identified a number of novel interactions for *influenza A* virus protein M1 with several human chaperones including *HSPA1A*, which is known to serve as a post-attachment receptor for rotavirus A entry into the cell [92].

We followed up our predicted host-pathogen interactions by analyzing the PPI profiles for *HIV-1* and *HIV-2*. In case of human infectivity, *HIV-1* is known to be a more virulent strain and therefore it is reasonable to assume that PPI profiles for the two viruses may reflect this difference. Indeed, we found several high confidence interactions specific to *HIV-1*. For example, *HIV-1* virion infectivity factor *Vif* forms novel interactions with *APOBEC3A*, *APOBEC3B*, and *APOBEC3D* that belong to the family of APOBEC deaminase proteins. Members of *APOBEC* family, for example *APOBEC3G*, are known to form an innate resistance to *HIV* infection and

hence they are deactivated by the virus during viral infection (Isolation of a human gene that inhibits *HIV-1* infection and is suppressed by the viral *Vif* protein [111]. Very interestingly, in agreement with our observation, *APOBEC3G* is reported to be deactivated by *HIV-1* protein *Vif* in a species specific manner [76].

6.5 Summary

PIPE3, the latest version of the PIPE algorithm, demonstrates how the PIPE algorithm can be applied to other organisms such as *C. elegans, S. pombe, E.coli* but most importantly *H. sapiens* (human). This expanded version is shown to successfully predict PPIs in other organisms but is also shown to be able to predict PPIs in organisms using other organisms as sources of information. The important implication of this feature is the ability to predict PPIs in new or unstudied organisms even given the lack of known interactions. PIPE3 has been used to do a genome–wide scan of the *S. pombe* organism as well as new PPI predictions in *H. sapiens, Hepatitis C, Influenza A, HIV-1* and *HIV-2.*

Chapter 7

PIPE Web Portal

7.1 Introduction

PIPE has been proven a useful tool to detect novel PPIs first in *S. cerevisiae* then in many other organisms. However a tool is not very useful if no one can use it. In order to give users access to the PIPE algorithm we created a web-portal for every version of PIPE to date: PIPE (Version 1), PIPE2 and PIPE3. In this chapter we will discuss each web–portal version and how they differ from each other.

7.2 PIPE Portal

The original PIPE web–portal (Figure 7.1) has been online since 2006 and has been used by hundreds of researchers across the world. Since the initial version of PIPE was very computationally intensive it offered researchers a portal to run their PPI predictions for *S. cerevisiae* without needing their own computer server.

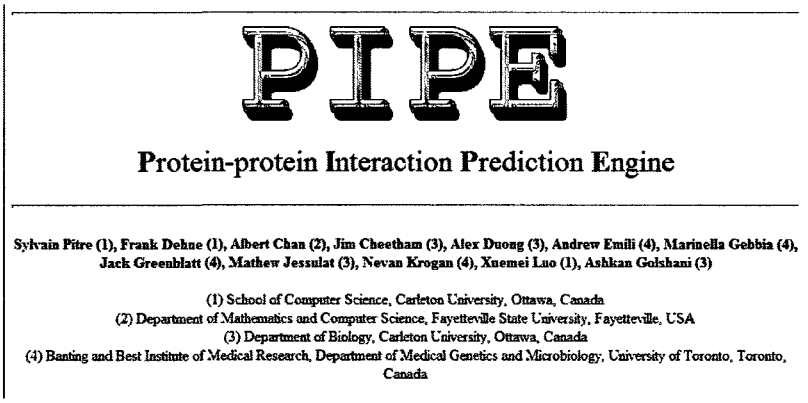

Figure 7.1: PIPE web–portal (Version 1) found at http://cgmlab.carleton.ca/PIPE/.

The settings included in this first version of the web–portal are fairly limited

(Figure 7.2). The user must enter two protein sequences along with identifying names on which PIPE will run. Due to the runtime of the original PIPE, both sequences are limited to 500AA each. After running the protein pair the portal will present the result matrix as well as a 3D graph representation. The interpretation of the graph is left to the user. The user can input any protein sequence (from any organism) however this version does not offer the settings necessary to adjust the algorithm for other organisms. Also, the database used is strictly composed of known *S. cerevisiae* PPIs.

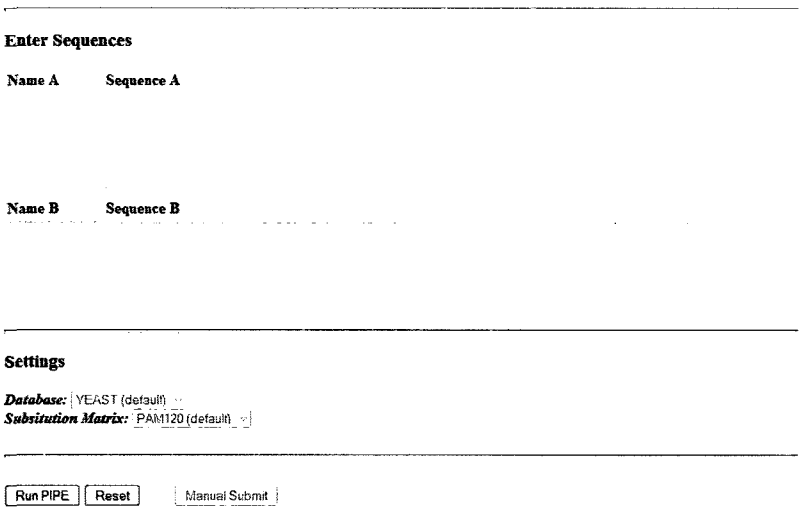

Figure 7.2: Settings available on the PIPE (Version 1) web–portal. The user can enter two protein names and sequences for prediction. The output includes the matrix H.

7.3 PIPE2 Portal

The second version of the PIPE web–portal (PIPE2, Figure 7.3) has been online since 2008. It has been highly–accessed in its 1.5 years of availability. The largest improvement to the original PIPE portal is in the prediction runtime. While the original PIPE algorithm could take several hours to make a prediction, PIPE2 has an average runtime of under 1/2 second. This speedup finally offered the user the possibility to run dozens of interactions in a short amount of time.

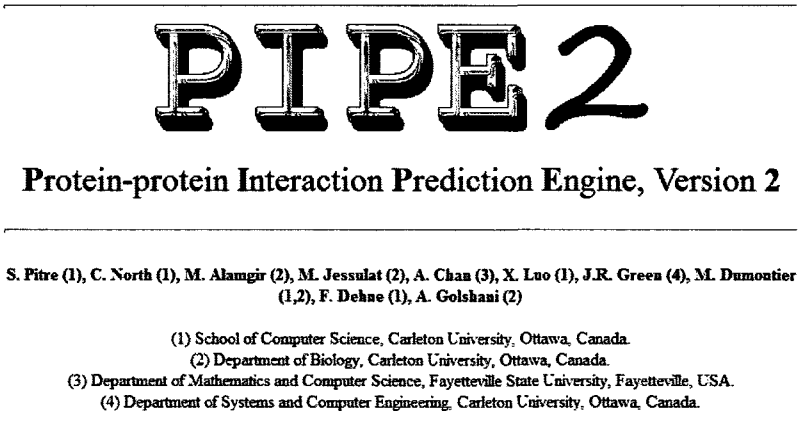

Figure 7.3: PIPE2 web–portal: http://pipe.cgmlab.org/

Since the PIPE2 algorithm is based on a pre-computed database, the user must chose two *S. cerevisiae* proteins from pre-defined lists (Figure 7.4). The **Cutoff Value** used to identify positive interactions can be changed to other settings. For example an 89% specificity might be acceptable if a user wishes to run a small number of pairs, but a higher specificity should be used at the expense of lower sensitivity for large number of pairs. After running the algorithm on the query proteins the results page will display some statistics and allow the user to download the result matrix as well its 3D representation.

7.4 PIPE3 Portal

The most recent version of the PIPE portal, PIPE3 (Figure 7.5), offers an important option not found in previous versions: multiple organism databases. Users now have the option to make PPI predictions using several pre-computed database (Figure 7.6) which includes: *S. cerevisiae*, *H. sapiens*, *E. coli*, *C. elegans*, *S. pombe* and the union of all listed organism (All).

Since this version uses pre-computed databases, the user must enter two valid protein names and must be from one of the organisms listed. However a user can run pairs from one organism while using another as the database (cross-organism predictions).

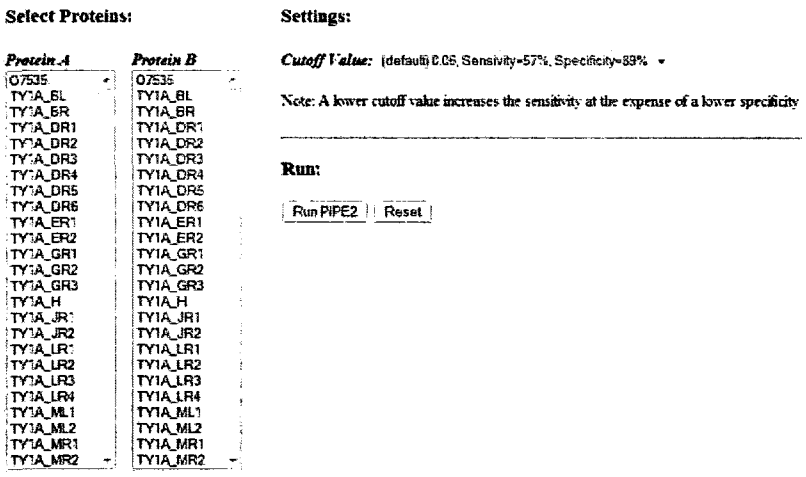

Figure 7.4: Settings available on the PIPE2 web–portal. The user must select two proteins from the pre-defined lists. The output offers statistics as well as answers if there is/is not an interaction between the query proteins.

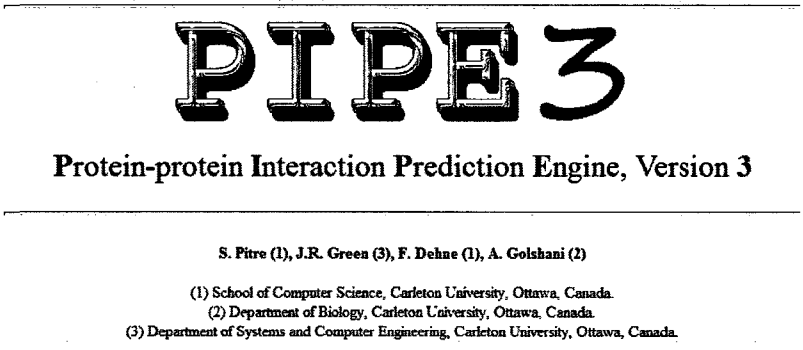

Figure 7.5: PIPE3 web–portal: http://cgmlab.carleton.ca/PIPE3/

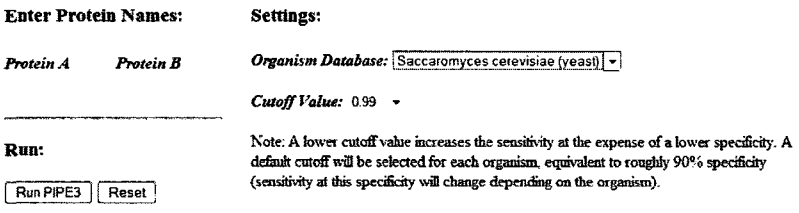

Figure 7.6: Settings available on the PIPE3 web–portal. A user can enter the names of both query protein (SwissProt ID), select the organism database used and the cutoff value used to decide if there is/is not an interaction.

Chapter 8

Conclusion

8.1 Introduction

In this chapter we will summarize the contributions made so far in this thesis in the field of protein–protein interaction prediction as well as summarize the future work proposed.

8.2 Summary of Contributions

This thesis presented a new algorithm for protein–protein interaction based on the re-occurring short polypeptide sequences between known interacting protein pairs called PIPE: **P**rotein–protein **I**nteraction **P**rediction **E**ngine [94]. This algorithm makes it possible to detect protein–protein interactions in *S. cerevisiae* using only sequence information (primary structure) as opposed to methods that rely on physical structure or known motifs or domains. Also this computational method is not restricted to the limitations of *in vivo* testing. It has been used to predict novel interactions and a novel process for which the internal structure has been confirmed by tandem affinity purification (TAP tag)[94].

An improved version of the algorithm called PIPE2 was also presented [95]. PIPE2 brings drastic speed improvement over PIPE along with higher specificity at the expense of lower sensitivity. The increased speedup of is due to optimized window comparisons, pre-computation and query approach. A modified version of the median filter is used on the results to improve the specificity which decreases the expected number of false positives significantly. However the filter has the negative but expected side–effect of reducing the sensitivity. With the increased speed it was possible to run all possible combinations of protein pairs in the *S. cerevisiae* genome

($\approx$ 20 million pairs). The PIPE2 interaction results list is larger than recent large–scale experiments using traditional methods [95].

Finally, the latest version of the algorithm named PIPE3 demonstrates how the PIPE algorithm can be applied to other organisms such as *C. elegans, S. pombe, E.coli* but most importantly *H. sapiens* (human). This expanded version is shown to successfully predict PPIs in other organisms but is also shown to be able to predict PPIs in organisms using other organisms as sources of information. The important implication of this feature is the ability to predict PPIs in new or unstudied organisms even given the lack of known interactions. PIPE3 has been used to do a genome–wide scan of the *S. pombe* organism as well as new PPI predictions in *H. sapiens, Hepatitis C, Influenza A, HIV-1* and *HIV-2.*

The PIPE method and the improvements achieved in PIPE2 and PIPE3 makes this project a great tool for biologists to predict protein–protein interactions without having to use traditional methods and while only needing the primary structure of proteins as input.

8.3 Future Work

8.3.1 Improvements In Protein Scanning

The current PIPE algorithm uses a single sliding window of size w in each query proteins in order to predict interactions (Figure 8.1(a)). One might argue that this is simplifying too much the interactions between two proteins. We have demonstrated that even using this simplified algorithm can yield meaningful results but we know that in reality interactions between two proteins could involve many fragments on each protein.

- **Two sliding windows of fixed size with fixed gap**: The first possible improvement we propose to use two sliding windows of size w and allow a gap g between those two windows (Figure 8.1(b)). This would make the test more stringent: in order for a protein in the database to match query proteins A or B they would need to contain those two same fragments within g amino acids of each other.

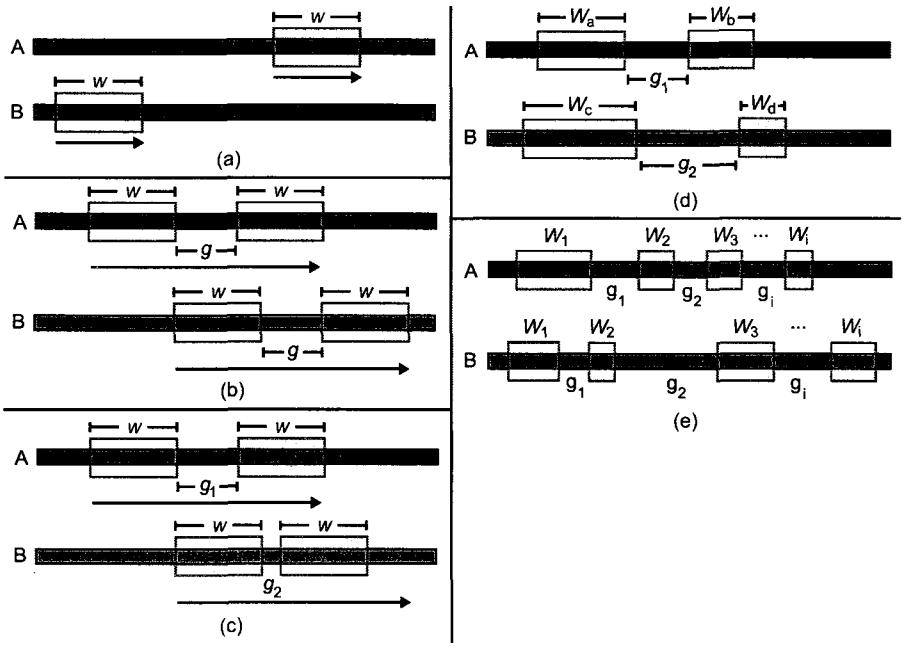

Figure 8.1: Different protein scanning techniques. (a) The default PIPE scanning technique (one sliding window of fixed size in each protein). (b) Two sliding windows of fixed size with a fixed gap. (c) Two sliding windows of fixed size and varying gap. (d) Two sliding windows of varying size and varying gap. (e) Multiple sliding windows of varying size with varying gaps.

- **Two sliding windows of fixed size with varying gap**: This is a variation of the previous modification. This improvement still uses two windows of fixed size but with a gap size of variable size (0 to g) between the windows in each query proteins (Figure 8.1(c)). The gap size can be different in each protein. This would allow, for example, two fragments of size 20 in one protein to be matched with a single piece of 40 in the other protein.

- **Two sliding windows of varying size with varying gap**: The next logical step in improving the protein scanning is to allow for variable sliding window size and gap in each protein (Figure 8.1(d)). This method would make it possible to match a larger fragment in protein A with a smaller piece in protein B as long as the fragment score is still above the defined threshold.

- **Multiple sliding windows of varying size with varying gaps**: The final

and most complex modification would allow multiple sliding windows of varying size along with varying gaps (Figure 8.1(e)). This method would produce a significant increase in complexity since we would be trying all possible combinations of fragments and gaps. One way to narrow down the search would be to bound the fragment sizes to a specified range (ex: 5-50 amino acids).

8.3.2 Other PIPE Projects

The PIPE project has grown from a PPI prediction engine to include other predictions. Most computational and traditional PPI methods limit themselves to predicting if an interaction exists between proteins. While it is possible to find the sites of interaction this is rarely done due to lengthy and costly experiments (i.e. by deletion experiments). Computational PPI methods are mostly trained on the identification of interactions rather than the sites of interaction. Most methods based on SVM for example will simply classify a protein pair as interacting or non-interacting. While the original PIPE was shown to be able to detect the sites of interaction in both proteins, that feature was never extended in PIPE2 or PIPE3. However, a new feature by Adam Amos-Binks accurately predict the interaction sites and present them to the user. This feature will be made available to PIPE users through an online web–portal.

PIPE3 has already been shown to be a good predictor of *H. sapiens* PPIs. An experiment we were anxious to start is a genome–wide scan of *H. sapiens*, however as we have seen this is simply not possible given the expected runtime (3 years using a 256-CPU compute cluster). Work is being done by Andrew Schoenrock to more effectively parallelize the work of such an experiment on a larger computer cluster. This would represent a huge achievement by the PIPE algorithm as it would stand as the first and only method capable of running genome–wide PPI prediction in *H. sapiens*.

8.4 Conclusions

In *Chapter 1*, we explained the need for a computational method to predict protein–protein interactions due to the drawbacks of traditional *in vivo* methods. We also shown computational methods which rely on unavailable data such as 3D structures or other information such as known domains. It was also shown in this thesis that we can predict protein–protein interactions in *S. cerevisiae* with a method based on the re-occurring short polypeptide sequences between known interacting protein pairs. The PIPE [94] and its subsequent versions PIPE2 [95] and PIPE3 have been shown to successfully detect meaningful novel interactions and complexes which have been in some cases confirmed by traditional experiments. In conclusion, we believe that PIPE is a great computational tool and has the potential to predict interactions in many other organisms than those presented with good accuracy.

Bibliography

[1] M. D. Adams et al. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–95, 2000.

[2] N. P. Allen, S. S. Patel, L. Huang, R. J. Chalkley, A. Burlingame, M. Lutzmann, E. C. Hurt, and M. Rexach. Deciphering networks of protein interactions at the nuclear pore complex. *Mol Cell Proteomics*, 1(12):930–46, 2002.

[3] E. Alm and A. P. Arkin. Biological networks. *Curr Opin Struct Biol*, 13(2):193–202, 2003.

[4] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–901, 2002.

[5] P. Aloy and R. B. Russell. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–2, 2003.

[6] E. Andres Leon, I. Ezkurdia, B. Garcia, A. Valencia, and D. Juan. EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res*, 37(Database issue):D629–35, 2009.

[7] M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H. C. Huang, A. Hirai, K. Tsuzuki, S. Nakamura, M. Altaf-Ul-Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, M. Kitagawa, M. Tomita, S. Kanaya, C. Wada, and H. Mori. Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res*, 16(5):686–91, 2006.

[8] M. Arnold, A. Nath, J. Hauber, and R. H. Kehlenbach. Multiple importins function as nuclear transport receptors for the Rev protein of human immunodeficiency virus type 1. *J Biol Chem*, 281(30):20883–90, 2006.

[9] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue. BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res*, 29(1):242–5, 2001.

[10] G. D. Bader, A. Heilbut, B. Andrews, M. Tyers, T. Hughes, and C. Boone. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends Cell Biol*, 13(7):344–56, 2003.

[11] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–7, 2002.

[12] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:i38–46, 2005.

[13] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2, 2006.

[14] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, 35(Database issue):D301–3, 2007.

[15] D. Betel, K. E. Breitkreuz, R. Isserlin, D. Dewar-Darch, M. Tyers, and C. W. Hogue. Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol*, 3(9):1783–9, 2007.

[16] J. R. Bock and D. A. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–60, 2001.

[17] C. T. Brown, Y. Xie, E. H. Davidson, and R. A. Cameron. Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. *BMC Bioinformatics*, 6:70, 2005.

[18] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–82, 2005.

[19] G. Butland, N. J. Krogan, J. Xu, W. H. Yang, H. Aoki, J. S. Li, N. Krogan, J. Menendez, G. Cagney, G. C. Kiani, M. G. Jessulat, N. Datta, I. Ivanov, M. G. Abouhaidar, A. Emili, J. Greenblatt, M. C. Ganoza, and A. Golshani. Investigating the in vivo activity of the DeaD protein using protein-protein interactions and the translational activity of structured chloramphenicol acetyltransferase mRNAs. *J Cell Biochem*, 100(3):642–52, 2007.

[20] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–4, 2007.

[21] K. R. Christie et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*, 32(Database issue):D311–4, 2004.

[22] J. M. Claverie. Gene number. What if there are only 30,000 human genes? *Science*, 291(5507):1255–7, 2001.

[23] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*, 6(3):439–50, 2007.

[24] S. E. Critchlow and S. P. Jackson. DNA end-joining: from yeast to man. *Trends Biochem Sci*, 23(10):394–8, 1998.

[25] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–8, 1998.

[26] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–8, 2002.

[27] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*, 30(1):69–72, 2002.

[28] A. Dziembowski and B. Seraphin. Recent developments in the analysis of protein complexes. *FEBS Lett*, 556(1-3):1–6, 2004.

[29] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10):529–36, 2002.

[30] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

[31] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.

[32] J. Espadaler, O. Romero-Isart, R. M. Jackson, and B. Oliva. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, 21(16):3360–8, 2005.

[33] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6, 1989.

[34] G. Franzot and O. Carugo. Computational approaches to protein-protein interaction. *J Struct Funct Genomics*, 4(4):245–55, 2003.

[35] A. C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.

[36] A. C. Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

[37] H. Ge, Z. Liu, G. M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*, 29(4):482–6, 2001.

[38] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, 2003.

[39] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546, 563–7, 1996.

[40] J. Goll and P. Uetz. The elusive yeast interactome. *Genome Biol*, 7(6):223, 2006.

[41] C. W. Gourlay, H. Dewar, D. T. Warren, R. Costa, N. Satish, and K. R. Ayscough. An interaction between Sla1p and Sla2p plays a role in regulating actin dynamics and endocytosis in budding yeast. *J Cell Sci*, 116(Pt 12):2551–64, 2003.

[42] A. Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31(14):4157–61, 2003.

[43] J. Guo, X. Wu, D. Y. Zhang, and K. Lin. Genome-wide inference of protein interaction sites: lessons from the yeast high-quality negative protein-protein interaction dataset. *Nucleic Acids Res*, 36(6):2002–11, 2008.

[44] D. S. Han, H. S. Kim, W. H. Jang, S. D. Lee, and J. K. Suh. PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res*, 32(21):6312–20, 2004.

[45] D. S. Han, H. S. Kim, W. H. Jang, S. D. Lee, and J. K. Suh. PreSPI: design and implementation of protein-protein interaction prediction service system. *Genome Inform*, 15(2):171–80, 2004.

[46] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.

[47] T. R. Hazbun and S. Fields. Networking proteins in yeast. *Proc Natl Acad Sci U S A*, 98(8):4277–8, 2001.

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

[49] H. Huang, B. M. Jedynak, and J. S. Bader. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol*, 3(11):e214, 2007.

[50] T. W. Huang, A. C. Tien, W. S. Huang, Y. C. Lee, C. L. Peng, H. H. Tseng, C. Y. Kao, and C. Y. Huang. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–6, 2004.

[51] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.

[52] A. Jazayeri, A. D. McAinsh, and S. P. Jackson. Saccharomyces cerevisiae Sin3p facilitates DNA double-strand break repair. *Proc Natl Acad Sci U S A*, 101(6):1644–9, 2004.

[53] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001.

[54] M. Jessulat, M. Alamgir, H. Salsali, J. Greenblatt, J. Xu, and A. Golshani. Interacting proteins Rtt109 and Vps75 affect the efficiency of non-homologous end-joining in Saccharomyces cerevisiae. *Arch Biochem Biophys*, 2007.

[55] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–72, 2009.

[56] W. K. Kim, J. Park, and J. K. Suh. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform*, 13:42–50, 2002.

[57] H. Korkaya, S. Jameel, D. Gupta, S. Tyagi, R. Kumar, M. Zafrullah, M. Mazumdar, S. K. Lal, L. Xiaofang, D. Sehgal, S. R. Das, and D. Sahal. The ORF3 protein of hepatitis E virus binds to Src homology 3 domains and activates MAPK. *J Biol Chem*, 276(45):42389–400, 2001.

[58] N. J. Krogan, J. Dover, A. Wood, J. Schneider, J. Heidt, M. A. Boateng, K. Dean, O. W. Ryan, A. Golshani, M. Johnston, J. F. Greenblatt, and A. Shilatifard. The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell*, 11(3):721–9, 2003.

[59] N. J. Krogan, M. Kim, A. Tong, A. Golshani, G. Cagney, V. Canadien, D. P. Richards, B. K. Beattie, A. Emili, C. Boone, A. Shilatifard, S. Buratowski, and

J. Greenblatt. Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol*, 23(12):4207–18, 2003.

[60] N. J. Krogan, M. H. Lam, J. Fillingham, M. C. Keogh, M. Gebbia, J. Li, N. Datta, G. Cagney, S. Buratowski, A. Emili, and J. F. Greenblatt. Proteasome involvement in the repair of DNA double-strand breaks. *Mol Cell*, 16(6):1027–34, 2004.

[61] N. J. Krogan et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–43, 2006.

[62] B. Kruger and T. Dandekar. Bioinformatical approaches to detect and analyze protein interactions. *Methods Mol Biol*, 564:401–31, 2009.

[63] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[64] P. Legrain, J. Wojcik, and J. M. Gauthier. Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet*, 17(6):346–52, 2001.

[65] W. P. Lehrach, D. Husmeier, and C. K. Williams. A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics*, 22(5):532–40, 2006.

[66] B. Li, S. G. Pattenden, D. Lee, J. Gutierrez, J. Chen, C. Seidel, J. Gerton, and J. L. Workman. Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling. *Proc Natl Acad Sci U S A*, 102(51):18385–90, 2005.

[67] D. Li, W. Liu, Z. Liu, J. Wang, Q. Liu, Y. Zhu, and F. He. Princess, a protein interaction confidence evaluation system with multiple data sources. *Mol Cell Proteomics*, 7(6):1043–52, 2008.

[68] S. Li et al. A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657):540–3, 2004.

[69] Y. Liu, I. Kim, and H. Zhao. Protein interaction predictions from diverse sources. *Drug Discov Today*, 13(9-10):409–16, 2008.

[70] L. Lu, H. Lu, and J. Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–64, 2002.

[71] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, 100(10):5772–7, 2003.

[72] M. Mann and A. Pandey. Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem Sci*, 26(1):54–61, 2001.

[73] E. M. Marcotte. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol*, 10(3):359–65, 2000.

[74] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, 1999.

[75] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–6, 1999.

[76] R. Mariani, D. Chen, B. Schrofelbauer, F. Navarro, R. Konig, B. Bollman, C. Munk, H. Nymark-McMahon, and N. R. Landau. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell*, 114(1):21–31, 2003.

[77] S. Martin, D. Roe, and J. L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21(2):218–26, 2005.

[78] N. Memarian, M. Jessulat, J. Alirezaie, N. Mir-Rashed, J. Xu, M. Zareie, M. Smith, and A. Golshani. Colony size measurement of the yeast gene deletion strains for functional genomics. *BMC Bioinformatics*, 8:117, 2007.

[79] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, 2002.

[80] J. Miller and I. Stagljar. Using the yeast two-hybrid system to identify interacting proteins. *Methods Mol Biol*, 261:247–62, 2004.

[81] N. Mizushima, T. Noda, and Y. Ohsumi. Apg16p is required for the function of the Apg12p-Apg5p conjugate in the yeast autophagy pathway. *EMBO J*, 18(14):3888–96, 1999.

[82] H. S. Najafabadi and R. Salavati. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol*, 9(5):R87, 2008.

[83] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12):e405, 2005.

[84] U. Oberholzer and M. A. Collart. Characterization of NOT5 that encodes a new component of the Not protein complex. *Gene*, 207(1):61–9, 1998.

[85] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy. PRISM: protein interactions by structural matching. *Nucleic Acids Res*, 33(Web Server issue):W331–6, 2005.

[86] J. Palecek, J. Hasek, and H. Ruis. Rpg1p/Tif32p, a subunit of translation initiation factor 3, interacts with actin-associated protein Sla2p. *Biochem Biophys Res Commun*, 282(5):1244–50, 2001.

[87] X. Pan, P. Ye, D. S. Yuan, X. Wang, J. S. Bader, and J. D. Boeke. A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell*, 124(5):1069–81, 2006.

[88] A. Pandey and M. Mann. Proteomics to study genes and genomes. *Nature*, 405(6788):837–46, 2000.

[89] Y. Park. Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. *BMC Bioinformatics*, 10(1):419, 2009.

[90] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–14, 2001.

[91] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8, 1999.

[92] J. Perez-Vargas, P. Romero, S. Lopez, and C. F. Arias. The peptide-binding and ATPase domains of recombinant hsc70 are required to interact with rotavirus and reduce its infectivity. *J Virol*, 80(7):3322–31, 2006.

[93] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani. Computational methods for predicting protein-protein interactions. *Adv Biochem Eng Biotechnol*, 2008.

[94] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, 7:365, 2006.

[95] S. Pitre, C. North, M. Alamgir, M. Jessulat, A. Chan, X. Luo, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani. Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res*, 36(13):4286–94, 2008.

[96] R. Poirey, L. Despons, V. Leh, M. J. Lafuente, S. Potier, J. L. Souciet, and J. C. Jauniaux. Functional analysis of the Saccharomyces cerevisiae DUP240 multigene family reveals membrane-associated proteins that are not essential for cell viability. *Microbiology*, 148(Pt 7):2111–23, 2002.

[97] N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*, 21(4):162–9, 2003.

[98] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res*, 37(3):825–31, 2009.

[99] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–29, 2001.

[100] J. Regelmann, T. Schule, F. S. Josupeit, J. Horak, M. Rose, K. D. Entian, M. Thumm, and D. H. Wolf. Catabolite degradation of fructose-1,6-bisphosphatase in the yeast Saccharomyces cerevisiae: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol Biol Cell*, 14(4):1652–63, 2003.

[101] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–2, 1999.

[102] J. S. Rohila, M. Chen, R. Cerny, and M. E. Fromm. Improved tandem affinity purification tag and methods for isolation of protein heterocomplexes from plants. *Plant J*, 38(1):172–81, 2004.

[103] G. M. Rubin. The draft sequences. Comparing species. *Nature*, 409(6822):820–1, 2001.

[104] V. Rubio, Y. Shen, Y. Saijo, Y. Liu, G. Gusmaroli, S. P. Dinesh-Kumar, and X. W. Deng. An alternative tandem affinity purification strategy applied to Arabidopsis protein complex isolation. *Plant J*, 41(5):767–78, 2005.

[105] L. Salwinski and D. Eisenberg. Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, 13(3):377–82, 2003.

[106] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–51, 2004.

[107] T. Sandmann, J. M. Herrmann, J. Dengjel, H. Schwarz, and A. Spang. Suppression of coatomer mutants by a new protein family with COPI and COPII

binding motifs in Saccharomyces cerevisiae. *Mol Biol Cell*, 14(8):3097–113, 2003.

[108] M. Schuldiner, S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, J. S. Weissman, and N. J. Krogan. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3):507–19, 2005.

[109] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–61, 2000.

[110] J. I. Semple, C. M. Sanderson, and R. D. Campbell. The jury is out on guilt by association trials. *Brief Funct Genomic Proteomic*, 1(1):40–52, 2002.

[111] A. M. Sheehy, N. C. Gaddis, J. D. Choi, and M. H. Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–50, 2002.

[112] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104(11):4337–41, 2007.

[113] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42, 2007.

[114] N. Simonis, J. F. Rual, A. R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, S. Cevik, N. Klitgord, C. Fan, P. Braun, N. Li, N. Ayivi-Guedehoussou, E. Dann, N. Bertin, D. Szeto, A. Dricot, M. A. Yildirim, C. Lin, A. S. de Smet, H. L. Kao, C. Simon, A. Smolyar, J. S. Ahn, M. Tewari, M. Boxem, S. Milstein, H. Yu, M. Dreze, J. Vandenhaute, K. C. Gunsalus, M. E. Cusick, D. E. Hill, J. Tavernier, F. P. Roth, and M. Vidal. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat Methods*, 6(1):47–54, 2009.

[115] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright. Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17, 2008.

[116] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–20, 1997.

[117] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–92, 2001.

[118] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–23, 2003.

[119] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–9, 2006.

[120] D. J. Stephens and G. Banting. The use of yeast two-hybrid screens in studies of protein:protein interactions involved in trafficking. *Traffic*, 1(10):763–8, 2000.

[121] H. Y. Tang, J. Xu, and M. Cai. Pan1p, End3p, and S1a1p, three yeast proteins required for normal cortical actin cytoskeleton organization, associate with each other and play essential roles in cell wall morphogenesis. *Mol Cell Biol*, 20(1):12–25, 2000.

[122] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–4, 2002.

[123] Y. Tsukamoto, J. Kato, and H. Ikeda. Silencing factors participate in DNA repair and recombination in Saccharomyces cerevisiae. *Nature*, 388(6645):900–3, 1997.

[124] C. L. Tucker, J. F. Gera, and P. Uetz. Towards an understanding of complex protein networks. *Trends Cell Biol*, 11(3):102–6, 2001.

[125] D. Tzamarias and K. Struhl. Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature*, 369(6483):758–61, 1994.

[126] D. Tzamarias and K. Struhl. Distinct TPR motifs of Cyc8 are involved in recruiting the Cyc8-Tup1 corepressor complex to differentially regulated promoters. *Genes Dev*, 9(7):821–31, 1995.

[127] P. Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–7, 2000.

[128] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[129] J. P. Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18 Suppl 1:S276–84, 2002.

[130] A. J. Walhout, S. J. Boulton, and M. Vidal. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast*, 17(2):88–94, 2000.

[131] H. Wang, E. Segal, A. Ben-Hur, Q. Li, M. Vidal, and D. Koller. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol*, 8(9):R192, 2007.

[132] R. H. Waterston et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

[133] P. J. Werler, E. Hartsuiker, and A. M. Carr. A simple Cre-loxP method for chromosomal N-terminal tagging of essential and non-essential Schizosaccharomyces pombe genes. *Gene*, 304:133–41, 2003.

[134] S. J. Wodak and R. Mendez. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol*, 14(2):242–9, 2004.

[135] A. Wood, N. J. Krogan, J. Dover, J. Schneider, J. Heidt, M. A. Boateng, K. Dean, A. Golshani, Y. Zhang, J. F. Greenblatt, M. Johnston, and A. Shilatifard. Brel, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. *Mol Cell*, 11(1):267–74, 2003.

[136] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–5, 2002.

[137] C. Yan, L. H. Lee, and L. I. Davis. Crm1p mediates regulated nuclear export of a yeast AP-1-like transcription factor. *Embo J*, 17(24):7416–29, 1998.

[138] T. Yoko-o, S. K. Roy, and Y. Jigami. Differences in in vivo acceptor specificity of two galactosyltransferases, the gmh3+ and gma12+ gene products from Schizosaccharomyces pombe. *Eur J Biochem*, 257(3):630–7, 1998.

[139] N. Zaki, S. Lazarova-Molnar, W. El-Hajj, and P. Campbell. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, 10:150, 2009.

[140] G. Zeng, X. Yu, and M. Cai. Regulation of yeast actin cytoskeleton-regulatory complex Pan1p/Sla1p/End3p by serine/threonine kinase Prk1p. *Mol Biol Cell*, 12(12):3759–72, 2001.

[141] K. X. Zhang and B. F. Ouellette. GAIA: a gram-based interaction analysis tool–an approach for identifying interacting domains in yeast. *BMC Bioinformatics*, 10 Suppl 1:S60, 2009.

[142] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–5, 2001.