Prediction and Analysis of Novel Interaction Motifs in Yeast

by

Yuan Gui

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

In

Biology

Carleton University
Ottawa, Ontario

© 2012, Yuan Gui

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: 978-0-494-93572-9

Our file Notre référence

ISBN: 978-0-494-93572-9

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 protègent 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

Proteins carry out function in the cell by forming stable and transient interactions with other proteins. These physical interactions are mediated by evolutionarily conserved sequences found within the binding sites of protein, and changes to these binding sites may lead to disruption of protein-protein interaction and onset of disease. In recent years bioinformatics tools have been developed to assist in identifying these sites of interactions. PIPE, or Protein-Protein Interaction Prediction Engine, is a computational tool that we have recently developed to predict protein interactions and locate the interacting sites by screening for short polypeptide sequences, and searching for homologous sequences in databases of known interactions. Using PIPE we have identified over 1000 interacting motif pairs in *Saccharomyces cerevisiae*. We have performed yeast two-hybrid analysis of several of these motifs and have identified possible sequences that may be responsible for mediating interaction.

Acknowledgements

I would like to thank everyone who has made this wonderful opportunity possible and a thoroughly enjoyable two years. I would like to express my deep and sincere gratitude to my supervisor Dr. Ashkan Golshani, for giving me the chance to undertake this research. His positivity, patience, and the keen ability to communicate complex ideas using the most clear and easy to understand examples is a skill that I will always remember. Without his knowledge and guidance this thesis would not have been possible.

I would like to thank all the past and present members of the Golshani lab for all their friendship and support, and for providing a positive atmosphere both inside and outside of the lab. I am especially indebted to Firoozeh Chalabian, for all of her hard work and assistance she provided me. Without her I would not have been able to complete my experiments in a timely manner. Lastly I would like to thank my parents for their love and constant support.

Table of Contents

Abstract.....	ii
Acknowledgements	iii
Table of Contents	iv
List of Abbreviations	vii
List of Tables	ix
List of Illustrations.....	x
List of Appendices.....	xi
1 Introduction.....	1
1.1 Protein-Protein Interactions.....	1
1.1.1 Types of Protein Interactions	2
1.1.2 Protein Domains and Motifs	5
1.2 Experimental Methods for Studying Protein-Protein Interactions	7
1.3 Computational Methods for Predicting Protein-Protein Interactions	11
1.4 PIPE.....	19
1.4.1 The PIPE Algorithm.....	19
1.5 Objective	23
2 Materials and Methods.....	25
2.1 PIPE Binding Site Prediction	25
2.2 Motif Selection	26
2.3 Yeast Two-Hybrid Vectors.....	27
2.4 Primer Design and Preparation.....	28
2.5 Extraction of S288c Genomic DNA	28
2.6 PCR of Motif Encoding Regions.....	31

2.7	Gel Electrophoresis for Confirmation of PCR	31
2.8	Gel Extraction of PCR Products	32
2.9	Digestion of Motif PCR Products.....	33
2.10	Purification of pBI880 and pBI881	33
2.11	Digestion of pBI880 and pBI881.....	34
2.12	Ligation of Motifs to pBI880 and pBI881	34
2.13	Transformation of Ligation to <i>E. coli</i>	35
2.14	Screening for Correct Clones Using Colony PCR.....	35
2.15	Screening for Correct Clones Using Restriction Digestion	36
2.16	Sequential Transformation of Single pBI880 and pBI881 Plasmids into MaV203	37
2.17	Transformation of Multiple pBI881 Plasmids into Yeast Containing pBI880 Plasmid.....	40
2.18	pBI881 Sequencing Primers	42
2.19	Spot Test on Selection Media for Yeast Two-Hybrid	43
2.20	β -galactosidase Assay.....	44
3	Results	46
3.1	PIPE Results	46
3.2	PCR and Cloning for Bait and Prey Plasmids	53
3.3	Sequencing Results.....	60
3.4	Yeast Two-hybrid Results	60
3.5	β -galactosidase Assay.....	66
4	Discussion.....	68
4.1	Analysis of Interacting Motifs and Parent Proteins.....	68
4.2	Amino Acid Sequence Complexity	77
4.3	Drug Discovery and Protein-Protein Interactions.....	80
4.4	Conclusion and Future Directions	83

Appendices.....	85
Appendix A Published Paper from Thesis.....	85
Appendix B Sequencing Results	86
B.1 Motif m2181	86
B.2 Motif m822	87
B.3 Motif m719	88
References.....	89

List of Abbreviations

3AT – 3-Amino-1,2,4-triazole

AD – Activating Domain

ADC1 – Alcohol Dehydrogenase 1

DBD – Binding Domain

DMSO – Dimethyl Sulfoxide

EDTA – Ethylenediaminetetraacetic acid

I2h – In Silico Two-Hybrid

IgG – Immunoglobulin G

LCR – Low Complexity Sequence

LiOAc – Lithium Acetate

MAP – Mitogen Activated Protein

MCS – Multiple Cloning Site

MS – Mass Spectrometry

NEB – New England Biolabs

OD – Optical Density

ONPG – Ortho-Nitrophenyl- β -galactoside

PAM – Percentage of Acceptable Point Mutations

PANTHER – Protein ANalysis THrough Evolutionary Relationships

PCA – Protein Complementation Assay

PCR – Polymerase Chain Reaction

PIPE – Protein-Protein Interaction Prediction Engine

PPI – Protein-Protein Interaction

PRD – Peptide Recognition Domain

RET – Resonance Energy Transfer

SGD – Saccharomyces Genome Database

TAP – Tandem Affinity Purification

TEV – Tobacco Etch Virus

UAS – Upstream Activation Sequence

Y2H – Yeast Two-Hybrid

YPD – Yeast Peptone Dextrose

List of Tables

Table 1: Examples of methods used for computational prediction of PPIs.....	13
Table 2: Total number of motifs generated by PIPE from the top 1000 PPI hits.....	47
Table 3: Representation of single amino acid repeats in PIPE results.....	49
Table 4: Single amino acid repeats in top 100 PIPE pairs from near-end list of motifs...	50
Table 5: Motifs selected for further analysis using Y2H system.....	51
Table 6: Pairs of motifs that are shown to interact based on Y2H analysis.....	68

List of Illustrations

Figure 1: Schematic of the Y2H system.	9
Figure 2: An illustration of the PIPE algorithm.....	22
Figure 3: PCR of motifs from S288c genomic DNA.....	54
Figure 4: Colony PCR for screening of correct motif into pBI880 and pBI881.....	57
Figure 5: Restriction digests of plasmids containing inserted motifs.....	59
Figure 6: Spot test for motifs m1068/m719.....	61
Figure 7: Spot tests for motifs m1068/m822 and m1068/m2181.....	62
Figure 8: Spot test for motifs m1955/m637 and m108/m2164.....	63
Figure 9: Spot test for motifs m1683/m536.....	64
Figure 10: Spot test for motifs m2188/m536.....	65
Figure 11: Relative β -galactosidase activity.....	67
Figure 12: Binding model of Yck2p and Yck3p to Sas10p.....	72
Figure 13: Hypothetical interaction model for Slt2p.....	74

List of Appendices

Appendix A: Published Paper from Thesis.....	85
Appendix B: Sequencing Results.....	86
B.1 Motif m2181.....	86
B.2 Motif m822.....	87
B.3 Motif m719.....	88

1 Introduction

1.1 Protein-Protein Interactions

Over the past several years it has become increasingly clear that the complexity of an organism is not reflective of the size of its genome. This is most apparent from the human genome, which is now known to be comprised of fewer protein-coding genes than originally anticipated (Lander et al., 2001). This is because the functions within cells are rarely determined by a single gene product, but rather due to thousands of gene products organised into a vast and complex network (Alberghina and Cirulli, 2010). It is now believed that the number and intricacy of the connections between individual gene products in this network is a major determinant of organism complexity (Xia et al., 2008). This network is called the interactome, and is comprised of an extensive number of both stable and transient protein-protein interactions (PPIs) among others (Sambourg and Thierry-Mieg, 2010). It is through these interactions that proteins carry out their function and participate in the majority of cellular processes, such as gene expression and signal transduction. Because of this fact, an understanding of biological mechanisms requires a systems level approach that goes beyond traditional studies of single components (Cusick et al., 2005). A better understanding of the interactions within a protein network can reveal the molecular mechanisms behind diseases (Kann, 2007).

For eukaryotes, one of the most useful model organisms for the study of interaction networks is the yeast *Saccharomyces cerevisiae*. The yeast genome was one of the first genomes to be sequenced and approximately 60% of its genes are conserved in humans. However, many of the functions of the discovered genes are still unknown. Out of the initial discovery of roughly 6000 gene products, almost a quarter are classified

as having unknown or unclear function (Peña-Castillo and Hughes, 2007). To date, there have been a number of large-scale studies performed to map the interaction network of *S. cerevisiae* and to determine the function of unknown proteins by determining their interacting partners. These include the study of binary interactions using large-scale yeast two-hybrid analysis (Ito et al., 2001; Uetz et al., 2000), large-scale protein fragment complementation assays (Tarassov et al., 2008), and the systematic analysis of protein complexes using tandem affinity purification followed by mass spectrometry (TAP-MSs) (Gavin et al., 2002).

While these large-scale studies provide a wealth of information regarding the binary interactions between proteins, they are unable to provide insight into the actual sites of interactions on proteins. Studying the interaction sites is not only important for quaternary structure prediction, but also for a comprehensive understanding of molecular processes and drug design (Ofra and Rost, 2003). Identification of the residues responsible for interaction may also simplify determining interacting proteins, as well as analysis of metabolic and signal transduction networks (Yan et al., 2004).

1.1.1 Types of Protein Interactions

The early study of protein interactions involved using the experimentally determined structures of proteins to identify key properties of protein interactions. The initial observation was that PPIs, in comparison to enzyme-ligand interactions, tend to involve large surface areas and lack cavities at their interaction sites (Jones and Thornton, 1996). Additionally, PPIs have a small group of residues that contribute the majority of the binding free energy necessary for interaction, surrounded by less important contact

residues. Only a small and complementary set of residues is necessary to maintain binding affinity (Clackson and Wells, 1995).

More recently, PPIs were loosely divided into two categories: globular protein interactions and domain-peptide interactions. Globular protein interactions are thought of as the canonical protein interaction, and are responsible for a diverse number of roles including cytoskeletal structure (actin, tubulin assemblies), macromolecular complexes (ribosome, RNA polymerase), and cellular signalling and regulatory pathways (receptor ligand interaction) (Hooda and Kim, 2012). They can be further categorized into homo-dimers (comprised of same subunit), obligate hetero-dimers (permanent complexes of two proteins), and hetero-dimers (transient domain-domain interactions). The general characteristics of globular protein interactions are a large interface surface area, high stability, and composed of two stable folded domains. The binding affinity covers a wide range, from micromolar to femtomolar. Conversely, domain-peptide interactions involve a folded domain binding to a linear stretch of amino acids. This folded domain is known as a peptide recognition domain (PRD), with several having already been discovered, including the SH3, PDZ, and WW domains. The SH3 domain, for example, specifically recognizes a core PXXP motif, where P is proline and X represents any amino acid. Not all PRDs interact exclusively with short amino acid motifs however. The PDZ domain normally binds to a ~4 amino acid motif at the extreme C-terminal ends of its binding partners, but has also been found to bind and form dimers with other PDZ domains, including itself through homodimerization (Nourry et al., 2003). Domain-peptide interactions are prevalent in cellular signalling and regulation pathways and a number of these domains have been studied and aggregated into databases such as DOMINO, a

database of known domain-peptide interactions (Ceol et al., 2007). Compared to globular protein interactions, the interaction surface is generally smaller, and binding affinity is from low-micromolar to nanomolar range.

PPIs can form both homo-oligomeric complexes and hetero-oligomeric complexes (Nooren and Thornton, 2003). Homo-oligomeric complexes are formed from interactions between two identical chains, can be formed in either an isologous or heterologous way. Isologous interactions involve the same surface on both monomers. These complexes can only further oligomerize using a different interface. Heterologous interactions on the other hand involve an interaction between two different interfaces. Aside from composition, the type of complex can be distinguished based on whether the complex is obligate or non-obligate. Obligate complexes cannot form a stable structure on their own *in vivo*, and require other components to form a stable complex. These complexes tend to also be functionally obligate. Non-obligate complexes on the other hand, are composed of proteins that can form stable structures *in vivo* on their own. Many of these proteins include intracellular signalling complexes, antibody-antigens, receptor-ligand, and enzyme-inhibitor complexes, and are often not co-localized initially, thus requiring independent stability.

Finally, proteins interactions can either form transient or permanent complexes. Permanent interaction are very stable and usually only exists in complex form, while a transient interactions is a short-lived association that dissociates rapidly. Structurally or functionally obligate interactions generally are permanent, while non-obligate interactions may be either permanent or transient (Nooren and Thornton, 2003).

Despite this number of characteristics of categorizing PPIs, many interactions do not fall into a specific archetype. The nature of an interaction often changes depending on the physiological conditions the proteins are subjected to. A continuum exists between obligate and non-obligate interactions, and the stability of complexes depends greatly on the cellular environment. A normally transient interaction can become permanent under the right conditions. It is also important to note that localization and function of a protein can often pre-determine the nature of interaction. Proteins involved in cellular signalling for example predominantly form transient interactions, as their functions require a steady and constant association and dissociation.

1.1.2 Protein Domains and Motifs

Definition of a domain varies depending on the context used. Domains are generally considered regions of compact protein structure, but they can either be categorized by protein function or structure. Structurally, domains that are from the same family share either significant sequence similarity or 3D structures (Bornberg-Bauer et al., 2004). In terms of function, experiments have shown that domains often represent functional units, and the total domain composition of a protein frequently determines its overall function (Copley et al., 2002). The human proteome contains several distinct interaction domains, present in hundreds of copies, that all serve to distinctly carry out different functions related to cellular organisation. For example, the human genome encodes 115 SH2 domains and 253 SH3 domains (Pawson and Nash, 2003). The SH2 domain typically binds a phosphorylated tyrosine residue, so is therefore responsible for tyrosine kinase signalling (Pawson and Scott, 1997). SH3 domains on the other hand, are

responsible for a wide variety of biological activities, including signal transduction, protein and vesicle trafficking, cytoskeletal organization, cell polarization, and organelle biogenesis (Mayer, 2001). The cell uses this limited number of domains in a variety of combinations as protein building blocks to direct the actions of regulatory systems. Because domain function is often related to the physical structure, and the physical structure is determined by the amino acid sequence, much of the current discovery of domain families focuses on the identification of conserved sequences. Homologous sequences are believed to have similar function, whatever the context, and these domain modules allow the transfer of functional information (Copley et al., 2001). It is believed that the higher the similarity in domain composition between two proteins, the more similar their functions (Hegyi and Gerstein, 2001). Well described domains have been aggregated in several databases, such as Pfam, which is a collection of protein domains and families, represented as multiple sequence alignments (Finn et al., 2008).

Protein motifs differ from domains in a few ways. While the terms tend to be used interchangeably, it is generally understood that domains are functional modules that can operate independently from a full-length protein. Motifs on the other hand represent relatively short amino acid sequences that are found within domains, and tend to be crucial for domain function. Motifs can either be categorized as functional motifs or sequence motifs. Sequence motifs, such as the zinc-finger motif consisting of two cysteines and two histidines (CXX(XX)CXXXXXXXXXXXXHXXXH) found in many DNA-binding proteins, can be identified primarily through the amino acid sequence, similar to sequence domains (Petsko and Ringe, 2004). Functional or structural motifs, such as the helix-turn-helix motif in many DNA-binding proteins, refer to a secondary

structure element that may have particular functional significance or define a portion of an independently folded domain. However, unlike structural domains, structural motifs cannot exist separately as a stable subunit if expressed outside of its native protein. Because of the importance to domain and ultimately protein function, motifs are generally more conserved.

As mentioned previously, domain-peptide interactions are commonly found in cellular signalling and regulation pathways. These pathways require the usage of modular protein domains to direct polypeptide association with each other as well other molecules, such as phospholipids, small molecules, or nucleic acids. These domains generally function by binding to a core group of amino acids, with flanking residues providing additional points of contact and allowing for greater selectivity. For example, the SH3 domain normally binds PXXP, but the SH3 domain in Csk can also bind to two additional C-terminal hydrophobic residues in PEP, increasing selectivity and specificity (Ghose et al., 2001). In some cases, the domain interaction with the core group of residues is sufficient for mediating interactions in cells.

1.2 Experimental Methods for Studying Protein-Protein Interactions

There currently exist many different techniques for studying PPIs. Two of the most commonly used ones are yeast two-hybrid (Y2H) and tandem affinity purification (TAP). Y2H, originally developed by Fields and Song in 1989, relies on the unique properties of eukaryotic transcription factors and their ability to function as modular units (Suter et al., 2008). The two subunits do not need to physically interact with each other to become active, and instead can function as long as the two are in close proximity. The

general method for studying interactions using Y2H is to clone the pair of genes for proteins being screened for interactions with each other, known as the bait and prey respectively, into a plasmid containing genes for the DNA binding domain (DBD) and the activation domain (AD) of the GAL4 transcription factor. The binding domain is responsible for binding to the upstream activating sequence of the reporter gene, and the activation domain binds to the binding domain. When bait and prey protein physically interact, the transcriptional factor is reconstituted by being in close proximity and becomes functional, resulting in the activation of a reporter gene and a measurable output (Figure 1).

Figure 1: Schematic of the Y2H system.

Proteins X and Y are fused to the binding domain (DBD) and activating domain (AD) of the GAL4 transcription factor, respectively. If there is no interaction (A), then the two subunits are separated from each other and there is no activation of the reporter. If an interaction does occur (B), then AD is recruited into proximity with DBD, activating transcription of the reporter gene (adapted from Brückner et al., 2009).

Commonly used reporters are growth in nutrient-deficient media, or production of quantifiable enzyme, such as β -galactosidase (Suter et al., 2008). Less uncommon reporters include fluorescent markers or reconstitution of essential signalling pathways. Additionally, in recent years there have been advances made to modify Y2H to study membrane proteins. This includes the use of split ubiquitin where two membrane proteins are fused with the halves of the ubiquitin protein (Iyer et al., 2005). One of the proteins is also tagged with a transcription factor attached to the ubiquitin half. When interacting proteins are brought together, the ubiquitin is targeted for degradation which releases the transcription factor from the cell membrane and can activate reporter genes. While commonly used to screen full-length protein, Y2H has also been shown to be useful in screening for interactions between small peptide fragments as small as 30 amino acids (Liao et al., 2006) (Yang et al., 1995). Y2H has been a tremendously useful tool in large-scale studies, such as the analysis of the entire yeast protein interactome (Ito et al., 2001).

Unlike Y2H, which determines binary interactions between two proteins, TAP tagging can be used to study and characterize protein complexes. TAP tagging is a form of affinity purification that relies on the use of two separate tags to purify interacting proteins. The protein of interest is fused to the TAP tag, which consists of the IgG binding domains of *Staphylococcus aureus* protein A (ProtA) and a calmodulin binding peptide, separated by a TEV protease cleavage site (Puig et al., 2001). The fusion protein binds to beads coated with IgG at its ProtA tag, and by association all proteins that physically interact with the target protein are also captured. Following cleavage of the tag at the TEV site which releases the protein, a secondary incubation with calmodulin

coated beads allows for the removal of remaining contaminants and TEV protease. The subsequent elution step to remove the captured protein complex is mild, and allows for retention of protein structure and secondary complexes. Interacting proteins can subsequently be analyzed using mass spectrometry or other methods of protein characterization. The advantage of TAP tagging is the production of the target protein is usually kept to the cell's natural levels to prevent any non-natural associations.

Other methods for studying PPIs include protein-fragment complementation assays (PCAs) and resonance-energy transfer systems (RET). While these methods are different, they all use the same underlying principle; each interacting protein is fused with one half of a detection molecule, and when brought together reconstitutes the fused molecule and creates a detectable signal. For PCA this can be two halves of a fluorescent protein or an enzyme that creates a colourimetric reaction. In the case of RET, bait and prey proteins are fused to acceptor and donor fluorescent molecules with different excitation wavelengths. Excitation of the donor creates energy transfer to the acceptor if bait and prey proteins are in close enough proximity, allowing for the study of spatial relationships (Petschnigg et al., 2010).

1.3 Computational Methods for Predicting Protein-Protein Interactions

Over the past several years bioinformatics and computational tools have been a great boon to the study of large-scale PPIs, and many different methods have been developed (Table 1). Predictive tools are derived from analysis of existing structure and sequence information, generated from experimental techniques, which have been curated into a large number of large publicly available databases (Plewczyński and Ginalski,

2009). In general, computational prediction is comprised of two main areas: determining whether two proteins are likely to interact and understanding the mechanisms of these interactions, often by identifying the residues responsible for these interactions (Skrabanek et al., 2008).

PPI prediction tool	Method of prediction	URLs (if publicly available)
PreSPI	Domain combinations	(N/A)
POINT	Orthologous interactome	(N/A)
PRISM	Structural matching	http://prism.cccb.ku.edu.tr/prism/
InterPreTS	Tertiary structure	http://www.russelllab.org/cgi-bin/tools/interprets.pl
In-silico two-hybrid	Correlated mutations	(N/A)
PIPE	Primary sequence	http://cgmlab.carleton.ca/PIPE-Sites/

Table 1: Examples of methods used for computational prediction of PPIs.

Initially, prediction of PPIs was limited to proteins that had a previously known three dimensional structural data available. These methods focused on the structural context of proteins, as well as using different physical property of the interacting surfaces to identify geometrically complementary regions between two proteins (Pazo et al., 1997). However this may not be feasible in situations where crystal structures have not been resolved and structural approaches are more limited in scope as only a small fraction of proteins have accurate three dimensional structural in databases. In recent years, advancements in whole genome sequencing have provided a wealth of sequence information, and allowed the establishment of a genomic context for a given gene in a genome. Whether two proteins are likely to interact can now be examined from not only its physical and structural properties, but also from a genomic and sequence level.

The earliest methods for determining the sites of interactions between proteins used the chemical and physical properties of proteins at interaction interfaces. For example, there is a large presence of hydrophobic residues at the interaction sites compared to the rest of the protein. It has been found that binding is highly favoured at these highly hydrophobic areas (Young et al. 1994). Hydrophobic residues such as Leu, Ile, Val, Phe, Tyr, and Met are found in greater abundance compared to polar residues such as Lys, Asp, and Glu, which are underrepresented (Janin et al., 1988). It is also suggested that the binding energy of the two proteins derives from burying of the hydrophobic surface areas (Chothia and Janin, 1975). Jones and Thornton used hydrophobicity, as well as five other parameters (solvation potential, residue interface propensity, planarity, protrusion, and accessible surface area) to isolate “surface patches”, defined by them as accessible surface areas of the protein greater than 5% that are

measurably different compared to the rest of the protein. Of the 28 homo-dimers that were studied, the highest hydrophobicity was found at the binding interfaces between monomers (Jones and Thornton, 1997). However, of the 20 different hetero-complexes that were also examined, no such trend was found. Some were found to have highly hydrophobic patches at their interaction interfaces, while others were more polar. This was explained by the nature of the interacting proteins. Those that exist as constitutive multimers and part of a permanent interaction were hydrophobic, while non-constitutive complexes due to a transient interaction that can exist as independent structures are more hydrophilic and the interacting interfaces were found to be some of the most polar regions on the protein surface. While using the chemical and physical properties of the interacting surfaces may be useful for determining interaction interfaces, this study relied on a small hand-selected dataset and presented contradictory information. Others have also suggested that identifying surface patches may be a poor method for determining which residues are responsible for interactions. Since surface patches only take into account residues that are exposed to the surface, buried residues with long side chains may be overlooked. As well it has been noted that the contribution of free energy to binding is not distributed evenly across the surface; some residues identified as part of a surface patch may form important contacts while others form none at all (Ofran and Rost, 2003).

Increasingly, more studies are relying on sequence information instead of structural and chemical properties to predict PPIs and interaction sites. Studies involved in predicting sites of interaction rely on the fact that the regions responsible for binding are critical and evolutionarily conserved. Kini and Evans identified proline brackets as

distinct barriers that separate interaction sites from neighbouring segments and may bracket or protect the integrity and conformation of the site (Kini and Evans, 1995). They found that the frequency was 2.5 times higher than expected from random distribution, and using this method they predicted a fibrin polymerization site in the epitope of the A α -chain of fibrinogen (Kini and Evans, 1996).

A genomic approach for predicting PPIs and sites of interactions uses the principle of correlated mutations. This is under the belief that interacting proteins undergo a process of co-evolution (Pazos et al., 1997). Over time, monomers that interact with each other will experience amino acid substitutions at its interacting interface that helps stabilize the interaction, and these stabilising mutations will be favoured in natural selection (Bennett et al., 1995). A mutation in the sequence of one protein will frequently have a compensatory mutation in its interacting partner. Expanding the theory beyond oligomers formed from the same subunit, if the interactions between two different proteins are the same physical nature as those found within proteins, then their changes at the sequence level should also be similar, and therefore can be detected using multiple sequence alignments (Pazos et al., 1997). Using their algorithm, they were able to predict the interaction site between the N-terminal and C-terminal domains of Hsp70 heat shock protein, which is responsible for peptide binding and release. As well they were able to predict and correctly identify binding sites on haemoglobin to other proteins. However the limitation of their method was that it was only able to predict sites of interactions for proteins already known to interact. Additionally it was only accurate in families of proteins with more than 15 members, with the conditions that their sequences have to be well distributed with both distant and close homologues. Pazo and Valencia later

proposed a new approach called *in silico* two-hybrid (i2h) based on their previous method using correlated mutations, which used a modified algorithm that allowed it to predict pairs of unknown interactions as well as determine the exact residues of binding (Pazos and Valencia, 2002). Predictions are achieved by creating two alignments of the interacting proteins, and concatenating the alignments to form a single cross-family alignment. A position-specific matrix is constructed from this alignment representing the distances between the residues corresponding to all sequence combinations at that position and a correlation function is then applied to detect mutations that are correlated both across and within families. A score representing correlated sites that could potentially indicate protein interactions is then determined. Using the haemoglobin α/β dimer as their example again, they were able to predict the residues for interaction: 36, 52, 88, and 84 of the α -subunit, and 102, 57, 46, and 64 of the β -subunit. A separate test was performed with eight known pairs of interacting proteins mixed together with 244 possible protein pairs, and i2h analysis picked out seven of them among the high scoring pairs. However, the downside to *in silico* two-hybrid is the same as previous studies using correlated mutations. Only proteins with a large number of homologues can be used to generate good quality alignments, otherwise the overall accuracy is poor and can dramatically increase noise in the procedure (Skrabanek et al., 2008).

Ofran and Rost attempted to create a method to predict sites of interactions on the basis of sequence alone. Previously this was determined to not be feasible, as the majority of the studies analyzing interaction sites were focused exposed regions of the protein. The residues that are responsible for interactions were thought to be distributed non-consecutively in sequence. If the physical nature of the interfaces was what governs

interactions, then studying sequence alone to determine sites of interactions would be considered impossible. However, using the basis of sequence features, they were able to determine six different types of protein interfaces with distinct amino acid composition. These interfaces differed on the basis of functional or structural association between residues (Ofra and Rost, 2003). The six different interfaces that they named were intra-domain, interfaces within one single structural domain, domain-domain, interfaces between two domains within one amino acid chain, homo-obligomer, interface between two permanently interacting identical chains, homo-complex, interface between transiently interacting identical chains, hetero-obligomer, interface between two permanently interacting different chains, and hetero-complex, interface between transiently interacting different chains. The difference in amino acid composition was significant enough that when sampling 1000 residues out of a specific interface type from a dataset, it would most closely resemble the composition of a second random sampling of 1000 residues from that same interface type compared to any of the other five interface types 63-100% of the time. For example, 1000 residues pulled from a dataset of intra-domain interfaces were closer to another 1000 residues from that same interface type 75.9% of the cases (Ofra and Rost, 2003). Using this information they were able to devise a method to predict a specific interface using only sequence information. The goal was to be able to ascertain whether or not residues responsible for interaction could be found to be clustered together in sequence, and whether it would be possible to predict these sites of interaction. They specifically focused on transient interactions between non-identical amino acid chains, and used the Protein Data Bank as their dataset, as it provided the greatest amount of information regarding true experimentally determined

positive interactions as well as proteins that were known to not interact (true negatives). Using this database they were able to find 1134 chains in 333 complexes, with 59559 residues in total. A residue was only considered if it was part of a PPI and ≤ 6 angstroms from an atom of the other protein. It was found that in stretches of nine amino acids, the majority contained five or more contacting residues. Contacting residues have a unique composition, this suggests that it is possible to detect and predict protein interactions and sites of interactions from sequence alone.

1.4 PIPE

PIPE, or Protein-Protein Interaction Prediction Engine, is a computational tool that was previously developed in our laboratory for prediction of protein interactions. The rationale for PIPE is that a finite number of short polypeptide sequences, smaller than traditional domains, are able to mediate interactions between two proteins (Pitre et al., 2006). Protein interactions are primarily governed by the physical contact mediated by their structure. The locations at which two proteins physically interact are the binding site, and commonly contain greater amino acid residue conservation (Pawson and Nash, 2003). It is these short amino acid sequences that PIPE identifies in known databases of interacting protein pairs, and uses to predict novel interactions without additional information about the protein pairs in question.

1.4.1 The PIPE Algorithm

The PIPE algorithm relies on a database of known pairs of PPIs, determined using several complimentary methods. Each protein sequence is represented on a graph (G) as

a node, with lines connecting it to its interacting partners. The sequences for two unknown interacting proteins are fed into the program: protein A and protein B. The first step involves breaking up protein A into small windows of w amino acids in length. The size of w can be arbitrarily defined. However, a fragment too small will result in many random matches, while one too large will result in too few. Starting at the beginning of the sequence, the window is searched for in the database of known interactions, and sequences are matched based on PAM120 scores (S_{PAM}). PAM, or Percentage of Acceptable Point Mutations, is a scoring system used to determine whether two sequences are similar, based on the likelihood of amino acid substitutions. Certain amino acid substitutions are more likely to be “accepted” biologically speaking; for example a substitution from one hydrophobic residue to another would be less likely to disrupt protein function. These conservative substitutions would be assigned a higher score than a non-conservative substitution, such as from a charged to an uncharged amino acid. The appropriate PAM score threshold was determined through trial and error, and it was found that a score of 35 was the most selective for differentiating interacting and non-interacting pairs (Pitre et al., 2006). If a match is found based on the PAM parameters, all the interacting partners belonging to the protein containing the fragment is put into a list. The process is repeated as the window is moved one amino acid down the entire sequence of A, until every window has been screened. This process is then repeated with protein B, only instead of searching the entire database of known interactions, windows are matched against the list of interacting partners generated from screening A. Matches are placed into a result matrix, with each row representing matching sequences in A and columns matching sequences in B. Every time PIPE finds a matching fragment for

protein B, an increment of 1 is added to the final score in the matrix. The final matrix is then translated into a 3D surface, with the score represented as the elevation and peak height (Figure 2).

Figure 2: An illustration of the PIPE algorithm.

In step 1 the database of known interactions and sequences of A and B are fed into the program, which generates a graph G. Step 2 involves searching for sequences a_i of size w . When a match is found, the neighbours of the protein the match is found in based on graph G is added to list R. Step 3 repeats the process, searching for sequence b_j in the list of neighbours R. For every successful match the value in the result matrix increases. After all searches are performed, the result matrix is converted into a 3D surface, where peaks represent possible sites of interactions (Step 4) (Adapted from Pitre et al., 2006).

1.5 Objective

There are two primary objectives of this study.

- To at least one novel motif that is responsible for mediating interactions between known and unknown interacting partners.
- To validate PIPE predictions and the hypothesis that short polypeptide sequences can be the sole mediators of interactions between proteins.

The first objective of the study is to find novel motifs that may be responsible for mediating interactions. From what we currently know of protein interactions, interaction sites are highly conserved and in many cases proteins bind to very short sequence motifs. Although many of these sequence motifs have been discovered, a growing number of novel motifs are reported in literature suggesting that many more exist that are responsible for mediating protein interactions. In a preliminary study PIPE successfully predicted the interaction sites between proteins with known interactions (Pitre et al. 2007). In one instance, it was successfully able to determine the binding sites between YNL243W and YBL007C, two yeast actin binding proteins. In a larger scale test, PIPE was able to accurately determine the interactions sites of 265 experimentally confirmed yeast interactions. It was found that the number of predicted yeast binding sites within 10% of the experimentally determined site on DOMINO was six times greater than what would be expected by chance. Using PIPE we should therefore be able to determine novel binding sites and sequences that can mediate interaction. We would like to be able

to identify at least one novel motif that is responsible for the mediation of protein-protein binding.

Using PIPE we determined 958 sites of interaction from 14438 novel pairs of yeast proteins that did not overlap with any annotations in the InterPro database. However, PIPE is still only computational software reliant on the accuracy of experimental data. While computational analysis provides a solid evidence for the accuracy of PIPE, experimental validation is still required to verify predictions. Using experimental techniques such as the Y2H assay we can validate PIPE predictions of novel interaction sites.

The second objective supports the basis behind PIPE, which is the idea that short sequences can be the sole mediator of interaction. It is suggested that protein interactions are composed of a finite number of these short sequence motifs, and given a large enough experimental dataset PIPE can locate these sequences. If these sequences can be the sole mediators of interactions, then by creating fusion proteins they should be able to artificially induce interaction between protein subunits that don't normally spontaneously interact. Based on current understanding of interaction dynamics between proteins this may not hold true, as multiple sites working in combination may be necessary for full interaction. For the purpose of our study however, we will use the assumption that single sequences can mediate interactions.

2 Materials and Methods

2.1 PIPE Binding Site Prediction

Since its initial inception, PIPE has been modified to be able to extract the exact motif locations using the result matrix generated (Amos-Binks et al., 2011). PIPE uses databases of known interactions as a training set, which is a set of data used to discover potentially predictive relationships. Known interactions in this case were from BioGRID, a large database that curates all known PPIs from published sources and currently has 69993 interactions for *S. cerevisiae* (Stark et al., 2006). The DOMINO PPI database was used to provide known sites of interactions for interacting proteins and to validate site prediction. DOMINO extracts experimental data from a wide variety of published sources and different experimental methods (Ceol et al., 2007). The predicted sites of interactions were extracted by PIPE's algorithm, and filtered against these databases of known interactions and interaction sites. These were then filtered against Interpro, which is a freely available database online that integrates information about protein families, domains, and functional sites from a variety of different sources and individual databases (Hunter et al., 2012).

2.2 Motif Selection

PIPE was used to screen the top 1000 pairs of proteins predicted to interact. The program then extracted lists of motifs based on certain criteria. Initial screening of motifs was based on the following three:

1. Novel – At least one motif in interacting pair is novel
2. Appropriate length – No greater than 30 amino acids in length
3. Near end – Occurring in first or last 100 amino acids of protein sequence

The first condition was used because the objective was to find unknown motifs, and all selected interacting pairs had to contain at least one novel motif. The second condition of amino acid length was added as a convenience for downstream cloning, as the initial plan was to synthesize the sequences as oligonucleotides under 100 base pairs in length. However, the method for obtaining the motifs was later changed to using PCR, so the length criterion became a rough guideline or around 24-50 amino acids. The last condition was added under the assumption that cloning of the motifs into the two-hybrid vectors would more accurately reflect the position of the motif in the native protein, as motifs are cloned at the C-terminal ends of the bait and prey vectors. Using these guidelines PIPE created lists of motifs in eight permutations of the above parameters: one novel, one novel good length, one novel near end, one novel good length near end, two novel, two novel good length, two novel near end, two novel good length near end. Each pair of motif sequence was analyzed using BLAST and other databases such as SGD (Saccharomyces Genome Database) to find protein information.

To confirm the novelty of the motifs, protein systematic names were queried in UniProt to find the accession ID, and then searched through Interpro to see if motifs belonged to any known domains. Known domains were sorted based on their amino acid positions, and were matched to the amino acid positions given by PIPE. As well, motifs were run through BLAST with the low complexity filter turned on. The filter uses the SEG program developed by Wootton and Federhen, which identifies and removes sequences of low complexity below a predetermined threshold from the BLAST alignment (Wootton and Federhen, 1993).

2.3 Yeast Two-Hybrid Vectors

The two vectors containing the bait and prey plasmids respectively, were pBI880 and pBI881, provided for usage by Dr. Shelley Hepworth (Carleton University). The two are modifications of the pPC62 and pPC86 vectors, respectively, through the addition of FLAG epitope and enterokinase cleavage site upstream of the MCS (Chevray and Nathans, 1992). Vector pBI880 contains the GAL4 binding domain (DB) as well as a gene for the production of leucine (*LEU2*), allowing for selection on media lacking leucine, while vector pBI881 carries the GAL4 activating domain (AT) alongside a tryptophan gene (*TRP1*). Both vectors also have an ampicillin gene to allow for selection when transformed in *E. coli* and replication origins (*col E1* for bacteria, *CEN6/ARS* for yeast) for propagation in both organisms (Kohalmi et al., 1997). Both vectors also contain multiple cloning sites with a number of restriction enzyme sites. The ones used for cloning motifs were the NotI and Sall restriction sites. The expression of both GAL – DB and –TA is under the control of yeast alcohol dehydrogenase 1 gene (*ADC1*) and its

promoter and termination sequences which gives constitutive expression of the two fusion proteins (Bennetzen and Hall, 1982).

2.4 Primer Design and Preparation

Primers were designed based on the sequences of motif pairs predicted by PIPE. The amino acid sequences were run through Protein BLAST to find the corresponding nucleotide sequence of the parent protein in the *Saccharomyces cerevisiae* strain S288c (*MATa*, *SUC2*, *gal2*, *mal*, *mel*, *flo1*, *flo8-1*, *hap1*, *ho*, *bio1*, *bio6*). Strain S288c was selected as it is the most commonly used yeast strain, with its sequence fully known and stored on the Saccharomyces Genome Database. The amino acid position on the protein for each motif was provided by PIPE. Primers were designed corresponding to the first and last 21 amino acids of each motif, and the length was further adjusted to bring T_m within range of 50-60 °C. Since the motifs were to be cloned into bait pBI880 and prey pBI881 plasmids, restriction sites were added to the ends of each primer. A SalI site was added to the 5' end of the forward primer with seven random nucleotides added as the guard sequence, and a NotI site added to the reverse primer with six random nucleotides acting as the guard sequence. The primers were ordered from Integrated DNA Technologies. Primers were diluted to a stock concentration of 100 uM concentration using Milli-Q water, and then further diluted to a 20 uM working concentration.

2.5 Extraction of S288c Genomic DNA

As the motifs were part of native yeast proteins, motifs were obtained through performing PCR using S288c genomic DNA. The initial step was to extract sufficient

genomic DNA in high enough concentration to perform the PCRs. The following buffers were required:

- STET buffer – 8% sucrose, 50 mM Tris pH 8, 5% Triton X-100, 50 mM EDTA
- Ammonium acetate – 7.5 M, 28.9 g in 50 ml dH₂O

Yeast extract peptone dextrose (YPD) media was the standard media used to grow all non-transformed yeast strains. 1L of liquid YPD was made by adding 10 g yeast extract and 20 g of peptone to 800 ml of distilled H₂O (dH₂O) in a 1L bottle. In a separate 1L flask, 20 g of dextrose (D-glucose) was added to 200 ml of dH₂O. Separate vessels were used for the sugar and amino acids to prevent Maillard reaction from occurring during autoclaving. The bottle was capped loosely and the flask covered in aluminium foil, and both containers were autoclaved for 20 minutes on the liquid cycle. Following autoclaving the bottle and flask were allowed to cool to a safe temperature for handling and the sugar was poured into the bottle under a hood, and swirled to mix. The media was allowed to cool at room temperature, and then stored at 4 °C.

Genomic DNA was extracted from freezer stocks of S288c created prior in the lab. Yeast was inoculated from freezer stocks into 5 ml of YPD in 15 ml culture tubes, and grown for two days at 30 °C in an incubated shaker. The following day, the absorbance was measured using a spectrophotometer. 1 ml of culture was added to each cuvette, and the spectrophotometer was blanked using YPD media with wavelength set to 600 nm. The optical density (OD) value should be around 1.0. If lower, the cells were allowed to grow for longer until OD₆₀₀ of 1.0 was reached. The cells were transferred to

50 ml Falcon tubes, and spun down at 4000 RPM at 4 °C for 10 minutes using a benchtop centrifuge. The supernatant was then discarded. 100 µl of STET buffer was added and the solution was vortexed briefly. 0.3 g of 0.45 mm glass beads was added and the culture was vortexed for 5 minutes at room temperature. An additional 100 µl of STET was added and the culture briefly vortexed. The tubes were placed in 100 °C boiling water for 3 minutes, removed and cooled briefly on ice for 2 minutes. As much liquid as possible was transferred without removing the beads as well to a 1.5 ml tube. The beads were then discarded. The liquid was centrifuged at 4 °C for 10 minutes and 100 µl of the supernatant was transferred to a fresh 1.5 ml Eppendorf tube. 500 µl of 4 °C 7.5 M ammonium acetate was added and mixed briefly. The tubes were incubated at -20 °C for 1 hour. Following the hour, the tubes were centrifuged at 4 °C for 10 minutes to separate the protein, which was visible as a small white pellet. 200 µl of the supernatant was transferred to 400 µl of ice cold 100% ethanol in a separate tube, and mixed briefly. The mixture was centrifuged for 10 minutes at 4 °C and the supernatant was discarded. No visible pellet was detected, but the bottoms of the tubes were washed with 200 µl of 70% ethanol. The tubes were centrifuged for 5 minutes and the supernatant was discarded. The tubes were placed with caps open in an incubator at 30 °C to allow all the remaining liquid to evaporate. After drying, 20 µl of dH₂O was added to resuspend the genomic DNA. The absorbances of genomic DNA were measured using a NanoDrop 1000 by loading 1 µl of plasmid and Milli-Q water to blank.

2.6 PCR of Motif Encoding Regions

Motifs were synthesized by running PCRs. Initial test PCRs were done on a small scale using the following reaction mixture: 0.5 μ l template, 0.5 mM forward primer, 0.5 mM reverse primer, 2 μ l 10x Taq buffer, 0.6 μ l 50 mM $MgCl_2$, 0.4 μ l 10 mM dNTPs, 1 μ l Taq, and 14.5 μ l dH_2O for a total reaction volume of 20 μ l. Taq, $MgCl_2$, and Taq buffer were purchased from New England Biolabs (NEB). The dNTPs were purchased from Invitrogen, and prepared by combining 100 μ l each of 25 mM dATP, dTTP, dGTP, and dCTP with 600 μ l of Milli-Q water to make a stock concentration of 10 mM. A separate tube consisting of the mixture minus the template was used as a negative control. PCR mixtures were combined on ice, and run in a thermocycler using the following temperature and time conditions: 95 °C for 5 min, 95 °C for 1 min, 51 °C for 1 min, 72 °C for 1 min, and 72 °C for 2 min for 36 cycles. After PCR products were run on an agarose gel to confirm a product was being produced, full-scale PCRs were performed to generate a high enough concentration for the motif band to excise from a gel. Amounts that were used was as follows: 6.5 μ l template, 6.5 μ l forward primer, 6.5 μ l reverse primer, 26 μ l 10x Taq buffer, 7.8 μ l 50 mM $MgCl_2$, 5.2 μ l 10 mM dNTPs, 13 μ l Taq, and 188.5 μ l dH_2O to make a final amount of 260 μ l total reaction volume, which was then divided into 20 μ l aliquots in 13 PCR tubes.

2.7 Gel Electrophoresis for Confirmation of PCR

PCR products were run on 1.5% agarose gels. 0.6 g agarose was mixed with 45 ml TAE buffer in a 250 ml flask, and microwaved for 30 seconds three times, or until all

agarose was dissolved and no more particles were visible in solution. TAE buffer was made with the following recipes:

- 50x TAE buffer – 242 g Tris base, 57.1 ml glacial acetic acid, 100 ml of 0.5 M EDTA pH 8.0, brought up to 1 L with dH₂O
- 1x TAE buffer – 20 ml of 50x TAE buffer and dH₂O up to 1 L total

The melted agarose was allowed to cool to approximately 60 °C until it was safe to handle, and 0.5 µg/ml of ethidium bromide was added to the gel under a flow hood. The flask was vigorously swirled to ensure adequate mixing of the ethidium bromide with the gel. The melted gel was then poured into a gel casting mold and allowed to solidify. 4 µl of 6x loading buffer was added directly to the PCR mixture and mixed. 12 µl of each PCR was loaded into each well, and the gel was run at 120 volts for 1 hour. After 1 hour the gels were visualized under UV.

2.8 Gel Extraction of PCR Products

An agarose gel was made using the same amounts of agarose and TAE buffer as above. Using a standard comb, the prongs were taped to make larger wells to accommodate the larger PCR volume. The PCR products from the 13 tubes were combined in a single 1.5 ml tube, and 60 µl of loading buffer was added to the 260 µl mixture. The entire mixture was loaded onto the gel in roughly 160 µl volume per lane, and the gel was run for 1.5 hours to achieve full separate of the products from remaining primer band. Motif bands were excised from the gel using a razor blade on a UV box,

taking care not to cut too much excess agarose or other bands. Bands were cut quickly to prevent prolonged exposure of the DNA to UV that may cause damage to the inserts. Prior to cutting, the weight of each 2 ml tube being used was measured. The motif bands in gel were transferred into the pre-weighed 2 ml tube, and then weighed again to find the total weight. The weight of the gel was then calculated. The DNA was purified from the gel using a Qiagen gel extraction kit and their standard protocol and eluted with 60 µl of Milli-Q water. The concentration of purified PCR products was measured using a Nanodrop.

2.9 Digestion of Motif PCR Products

Following extraction from the gel, the motifs were digested to ensure compatible ends on the motif in preparation for ligation. Enzymes used for digestion were purchased from NEB. Digestion was performed using the following amounts: 5 µg purified PCR product, 5 µl of Buffer 3 from NEB, 0.25 µg/µl of BSA from NEB, 0.5 µl of NotI, 0.5 µl of Sall, and 23.5 µl of dH₂O to make a final reaction volume of 50 µl. Digestions were incubated at 37 °C for two hours. Following incubation the restriction enzymes were heat inactivated by incubating the reaction at 65 °C for 15 minutes.

2.10 Purification of pBI880 and pBI881

E. coli stocks carrying the pBI880 and pBI881 plasmids were inoculated from a freezer stock overnight and purified using the PureLink™ Quick Plasmid Miniprep Kit (Product # K2100-10) from Invitrogen following their standard protocol. Plasmids were

eluted using 60 µl of Milli-Q water instead of elution buffer. Concentrations of purified plasmids were measured using a Nanodrop.

2.11 Digestion of pBI880 and pBI881

Vectors pBI880 and pBI881 were linearized with restriction enzymes. The following amounts were used for each reaction: 20 µl pBI880 or pBI881, 5 µl of Buffer 3 from NEB, 0.25 µg/µl of BSA from NEB, 5 units NotI, 10 units of SalI, and 23.5 µl of dH₂O to make a total reaction volume of 50 µl. Digestions were incubated at 37 °C for two hours. Following digestion, the reaction was run on a 1% agarose gel to confirm complete digestion. If digestion was sufficient, the entire reaction was run on another 1% gel, and the linearized vector was excised and purified from the gel.

2.12 Ligation of Motifs to pBI880 and pBI881

Motifs were inserted into bait and prey vectors using standard ligation. To calculate the amount of insert used the following formula was applied:

$$\text{Insert Mass (ng)} = 6 \times (\text{Insert Length (bp)} / \text{Vector Length (bp)}) \times \text{Vector Mass (ng)}$$

To the calculated amount, the following amounts of other reagents were added: 2 µl pBI880/pBI881, 2 µl ligation buffer, 200 units ligase, and enough volume of dH₂O to bring the final reaction volume to 20 µl. Ligations were incubated at room temperature for 1 hour.

2.13 Transformation of Ligation to *E. coli*

Ligations were then transformed into DH5 α competent cells. 5 μ l of ligation reaction was added to 100 μ l of competent cells and placed on ice for 30 minutes. The cells were then heat-shocked for 1 minute by placing the tubes in a 42 °C heat block. Cells were removed from heat and placed back on ice for 1 minute. 250 μ l of LB media was added to each tube and incubated in a shaker for 1 hour at 37 °C. After 1 hour, tubes were removed and 100 μ l of cells were plated onto LB plates containing 100 μ g/ml ampicillin. After allowing liquid to dry, plates were incubated at 37 °C overnight.

LB plates were made by dissolving 12.5 g of LB powder and 5 g of agar in 500 ml of dH₂O in a 2 L flask. A stir bar was placed in the flask and the mixture was stirred until the agar was dissolved. The stir bar was then removed and the flask covered with aluminum foil. The entire mixture was autoclaved for 20 minutes using the liquid cycle. Following autoclaving the dissolved media was allowed to cool at room temperature until save to touch, roughly 40 °C, and 500 μ l of a 100 mg/ml stock solution of ampicillin was added and mixed by swirling the flask. The media was poured into round plates until they were about halfway full from the top. The following day plates were marked to indicate ampicillin had been added, bagged and stored at 4 °C.

2.14 Screening for Correct Clones Using Colony PCR

To confirm digestions, colony PCR was performed using the following conditions. The following PCR reaction was used: 0.5 mM forward primer, 0.5 mM reverse primer, 2 μ l 10x Taq buffer, 0.6 μ l 50 mM MgCl₂, 0.4 μ l 10 mM dNTPs, 1.25 units/50 μ l Taq, and 15 μ l dH₂O for 20 μ l total reaction volume. The primers selected

were the same primers used to generate the original motifs. The PCR tubes were filled with the 15 µl of dH₂O first. Random colonies were picked off the transformation plates using a pipette tip, and added to the dH₂O prior to addition of other components. Using the same tip, cells were streaked out onto LB plates to save the selected clones. Plates were incubated overnight at 37 °C. Following streaking the cells the rest of the PCR reagents were added to the reaction. For the negative control, no template was added. As well a second negative control of the empty pBI880 and pBI881 vectors were used. The colony PCR was then run at the following cycling conditions: 95 °C for 5 min, 95 °C for 1 min, 51 °C for 1 min, 72 °C for 1 min, 72 °C for 2 min for 36 cycles. The PCR reactions were then run on a 1.5% agarose gel to check for presence of motif band.

2.15 Screening for Correct Clones Using Restriction Digestion

To confirm colony PCR for the correct clones, colonies were picked off the streaked plates, and grown in 2 ml LB media overnight. Plasmids were extracted using a PureLink™ Quick Plasmid Miniprep Kit. Plasmids were subsequently digested to excise the motif using the following reaction conditions: 12 µl plasmid, 2 µl buffer 3, 0.25 µg/µl BSA, 5 units NotI, 10 units SalI, and 4.5 µl dH₂O for a total reaction volume of 20 µl. Because motifs were relatively small compared to the vectors they were excised from, a large amount of plasmid needed to be used to be able to properly visualize the motif bands. The digestion was incubated at 37 °C for 1 hour, followed by 15 minutes at 65 °C to inactivate the enzymes. The digested plasmids were run on a 1.5% agarose gel at 120 V for 1 hour and visualized with ethidium bromide and UV.

2.16 Sequential Transformation of Single pBI880 and pBI881 Plasmids into MaV203

The yeast strain used for Y2H was MaV203 (*MATa*, *leu2-3,112*, *trp1-901*, *his3Δ200*, *ade2-101*, *gal4Δ*, *gal80Δ*, *SPAL10::URA3*, *GAL1::lacZ*, *HIS3_{UAS}* *GAL1::HIS3@LYS2*, *can1^R*, *cyh2^R*). This strain is auxotrophic for leucine and tryptophan due to *leu2* and *trp1* mutations, which allows for selection of successful transformation of bait and prey vectors. Correct plasmids were transformed sequentially into MaV203 yeast strain. Additionally, the following reagents were required:

- 1 M lithium acetate (LiAc)
 - Add 10.2 g of LiAc·2H₂O (lithium acetate dehydrate) to 80 ml of dH₂O and adjust the pH to 7.5 with acetic acid
 - Adjust the volume to 100 ml with dH₂O
 - Autoclave for 20 minutes on liquid cycle
- 0.1 M lithium acetate
 - Dilute 10 ml of 1 M lithium acetate with 90 ml of dH₂O
 - Autoclave for 20 minutes on liquid cycle
- 50% polyethylene glycol
 - Add 100 g of polyethylene glycol and bring the volume to 200 ml with dH₂O
- DMSO

MaV203 was inoculated into 5 ml of YPD from a freezer stock and grown overnight at 30 °C in a shaker. The next morning, 4 ml of the overnight culture was inoculated into

40 ml of fresh pre-warmed YPD, and allowed to grow for 3-4 hours or until OD₆₀₀ was in the range of 0.6-1. Once culture had reached appropriate density, the cells were spun down at 3000 RPM for 2 minutes. The supernatant was discarded, and the pellets were vortexed with 10 ml of water. The cells were centrifuged again at 3000 RPM for 2 minutes, and the supernatant was discarded. 1 ml of 0.1 M lithium acetate (LiOAc) was added to the pellet, and the cells were gently vortexed. The cells were centrifuged again at 3000 RPM for 2 minutes, and the supernatant discarded. The pellet was then resuspended in 500 µl LiOAc, and 100 µl for each transformation as well as control of the solution was transferred to a clean 1.5 ml tube. For each transformation, the following was added to the 100 µl of cells: 100 ng of plasmid (volume calculated based on concentration of plasmid), 50 µl ssDNA, 240 µl 50% PEG, and 36 µl of 1 M LiOAc.

For the transformation of each plasmid, a negative control was used, which consisted of all the transformation reagents without the addition of plasmid. The transformation mixture was incubated at 30 °C for 40 minutes. After every 10 minutes during the incubation, the cells were removed from the incubator and inverted multiple times to ensure adequate mixing. After incubation, 35.5 µl of DMSO was added to each tube, and the tubes were inverted to mix. The tubes were then incubated at 42 °C for 20 minutes and then centrifuged at 8000 RPM for 2 minutes. The supernatants were discarded, and pellets resuspended in 110 µl of dH₂O. The entire resuspension was plated onto corresponding selection media depending on which plasmid was transformed first. If the first transformation was with a pBI880 vector, the cells were plated on a synthetic complete media lacking leucine (-Leu). If the first transformation was with a pBI881 vector however, the cells were plated on a synthetic complete media lacking

tryptophan (-Trp). Plates were incubated at 30 °C for two days or until colonies were visible. If colonies formed following the transformation of the first plasmid, a second transformation was carried out. The cells were inoculated from the first transformation plate into minimal liquid media of the same selection as the first transformation, and grown overnight. For example, if the first plasmid transformed was pBI880 and plated onto -Leu plates, colonies were picked off from it and inoculated into -Leu liquid. If the first plasmid was pBI881, then colonies were inoculated into -Trp. The protocol for the second transformation was identical to the first. For plating the transformation containing both plasmids, plates lacking both leucine and tryptophan were used.

The plates were made using a synthetic complete drop-out mixture of amino acids. To make the solid and liquid media, a drop-out powder mixture had to be made first. The mixture was made by combining 2 g of the following amino acid composition: adenine sulfate, alanine, arginine, aspartic acid, asparagine, cysteine, glutamic acid, glutamine, glycine, isoleucine, lysine, methionine, phenylalanine, proline, serine, threonine, tyrosine, and valine. The amino acids were added to a 250 ml flask covered with aluminum foil to prevent light exposure, and mixed vigorously using a vortex to ensure all amino acids were distributed evenly. The flask was then covered with a piece of aluminum foil and stored in a 4 °C fridge.

Selection plates were made in three separate flasks. In one 500 ml flask, 2.68 g yeast nitrogen base and 0.54 g of the amino acid drop-out mix prepared above were added to 100 ml of dH₂O. Depending on the selection media, 0.03 g of each amino acid not being selected was added back to the mixture. For example, for a -Leu plate, 0.03 g of histidine, uracil, and tryptophan were added back in. In a second 500 ml flask 8 g of

glucose were added to 100 ml of dH₂O. In a 1 L flask, 8 g agar was mixed with 200 ml dH₂O. All three flasks were covered with aluminum foil and autoclaved using the liquid cycle for 20 minutes. After autoclaving the flasks were allowed to cool to a safe temperature for handling and mixed together in the 1 L flask containing the melted agar. The flask was swirled vigorously to ensure the agar was evenly mixed, and the mixture poured into the plates and allowed to solidify overnight. The next day the plates were bagged and stored at 4 °C.

2.17 Transformation of Multiple pBI881 Plasmids into Yeast Containing pBI880 Plasmid

For one set of motifs that was predicted to interact with several different motifs, the second transformation was performed using a mixture of plasmids. After successful transformation of the first pBI880 plasmid, the second transformation was performed using the same method as above. However instead of using a single pBI881 plasmid, a mix of seven different plasmids was used instead. The mixture was made by combining 10 µl of each pBI881 plasmid containing a motif together. For the transformation, 200 ng of the combined plasmids was added to the competent yeast cells, and the rest of the transformation was carried out as usual. Transformations were plated onto media lacking –Leu –Trp solid media.

To screen these colonies and find out what second plasmid was taken up by the yeast, 30 colonies were picked and streaked onto –Leu –Trp –Ura plates. Plates were incubated at 30 °C and growth was observed. Colonies that were shown to grow on –Ura

media were selected for further examination and spot tests. Freezer stocks were also made with these cells, by adding 400 µl of 60% glycerol to 800 µl of cells.

To determine which motifs were shown to interact, it was necessary to be able to extract the single pBI881 plasmid containing the motif from the double transformed yeast. To do this, the pBI880 plasmid needed to be expelled from the yeast cells. The transformants that were shown to grow on –Leu –Trp –Ura solid media were picked and inoculated into 5 ml of liquid –Trp media. The cells were then incubated for five days until they had grown sufficiently. 20 µl of the cells were then pipetted onto solid –Trp media, and then incubated until colonies formed. These colonies were subsequently picked and inoculated back into 5 ml of –Trp liquid, and incubated for two days. This process of regrowing cells in –Trp media was repeated for five cycles. After the fifth cycle, cells from the liquid culture were plated on both single selection plates –Trp and –Leu and incubated at 30 °C until growth was observed. If growth was observed on the –Trp plate, but not the –Leu plate, than this was indication that the cells had successfully ejected the pBI880 plasmid. These cells were then picked and inoculated in –Leu liquid media so the plasmids could be purified from the yeast cells. Plasmid purification from yeast was performed using the E.Z.N.A. Yeast Plasmid Kit purchased from Omega Bio-Tek, following the standard protocol provided. The concentrations of plasmids obtained were measured using a Nanodrop. Following purification from yeast, the plasmids were re-transformed into DH5α *E. coli* competent cells to allow for replication to quantities needed for sequencing. The protocol for *E. coli* transformation follows the one listed previously, and transformants were plated onto LB media containing 100 µg/ml ampicillin, and incubated at 37 °C overnight. The following day, colonies from these

plates were then picked and inoculated into liquid LB media containing 100 µg/ml ampicillin, and incubated in a shaker at 37 °C overnight. The next day, the pBI881 plasmids were purified from the *E. coli* cells using Invitrogen's PureLink™ Quick Plasmid Miniprep Kit and eluted with 60 µl of Milli-Q water. The plasmid concentrations were measured using a Nanodrop.

2.18 pBI881 Sequencing Primers

Primers used for sequencing of pBI881 plasmids were designed based on its pPC86 predecessor, as the only modification was the addition of the FLAG epitope and a cleavage site. The full sequence of pPC86 is online on Invitrogen's website. The following primers were designed:

- pBI880 forward primer (GGACCAAACCTGCGTATAACGCGTT)
- pBI881 reverse primer (CTGGCAAGGTAGACAAGCCGACAA)

The forward primer binds to a sequence starting 159 base pairs upstream of the SalI cutting site, within the *GAL4* activating domain gene. The reverse primer binds 101 bp downstream from the NotI restriction site within the alcohol dehydrogenase terminator sequence. Both primers were ordered from Integrated DNA Technologies and made to a concentration of 20 mM. Purified plasmids and primers were sent to Bio Basic Inc. for sequencing.

2.19 Spot Test on Selection Media for Yeast Two-Hybrid

To determine whether interaction between motifs was occurring, a spot test onto selection media was performed. The selection plates used were SC –Leu – Trp –Ura, and SC –Leu – Trp –His +3AT. The plates were made using the same protocol as described previously. 3AT, or 3-Amino-1,2,4-triazole, was added at a concentration of 3 mM to –His plates to decrease base levels of histidine production. 3AT is a competitive inhibitor of the *HIS3* product, thereby only allowing cells expressing higher levels of histidine to survive.

Colonies were picked from plates containing double transformed yeast, and inoculated into –Leu –Trp liquid media, and incubated in a shaker for two days. The OD₆₀₀ of the cells were measured using a spectrophotometer and –Leu –Trp liquid media to blank the machine, to ensure that the absorbances were close to 1. An OD₆₀₀ of 1 is approximately 1×10^7 yeast cells/ml. If any samples were too high, the cells were diluted down accordingly using –Leu –Trp liquid media. For each colony picked, four dilutions of cells were used: 1×10^7 cells/ml, 1×10^6 cells/ml, 1×10^5 cells/ml, and 1×10^4 cells/ml. Dilutions were made by serially diluting the original sample into –Leu –Trp liquid media. 100 µl of the cells were added to 900 µl of the liquid media in a separate 1.5 ml Eppendorf tube. The tube was vortexed to ensure cells were thoroughly mixed, and 100 µl of the diluted cells were then added to another 900 µl of media. This process was repeated until all dilutions were made. 15 µl of each dilution was spotted onto the selection plates, in sequential fashion. The plates were kept open under a hood until the liquid had dried completely, and then incubated at 30 °C for two nights, or until colonies had started forming.

2.20 β -galactosidase Assay

As an additional method for determining whether motifs interact, β -galactosidase assays were carried out. The following buffers were necessary for the assay:

- Z-Buffer
 - 16.1 g $\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$
 - 5.5 g $\text{NaH}_2\text{PO}_4 \cdot 7\text{H}_2\text{O}$
 - 0.75 g KCl
 - 0.246 g $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$
 - The volume was brought to 1 L with dH_2O and the pH was adjusted to 7.0, and the buffer was autoclaved for 20 minutes on the liquid cycle.
- 1M Na_2CO_3
 - 5.3 g powder Na_2CO_3
 - Volume brought to 50 ml with dH_2O , and stored at room temperature
- 4 mg/ml 0-nitrophenol (ONPG)
 - 4 mg of ONPG to 1 ml of 0.1 M KH_2PO_4 and the pH was adjusted to 7.0
- 0.1% SDS
 - 0.1 g of SDS was dissolved in 100 ml dH_2O

Cells were inoculated and grown in 5 ml of –Leu –Trp media for two days. The absorbance of the overnight culture was measured to ensure the OD_{600} was greater than 1. Tubes with 5 ml of –Leu –Trp were prepared, and yeast cells were diluted in the media so the OD_{600} of each was brought down to approximately 0.2 to 0.3. To calculate the

volume of overnight culture needed for dilution, the following formula was used: $V_{\text{culture}} (\text{ul}) = (5000 \times 0.25) / (\text{OD}_{600} \text{ of overnight culture})$. The cells were then incubated at 30 °C in a 200 RPM shaker and allowed to grow until the OD_{600} of each reached 1, approximately 4 hours.

A tube of ONPG was removed from the freezer and placed in a 30 °C incubator to allow time to warm up. The cells were then transferred to 50 ml tubes, and centrifuged at 4000 RPM for 4 minutes at 20 °C. The supernatants were discarded and the cells were resuspended in 3 ml Z-buffer. The tubes were centrifuged again at 4000 RPM, and the supernatant discarded. The cells were then resuspended in 0.5 ml Z-buffer, vortexed, and kept on ice. In a separate 1.5 ml tube 200 μl of cells were added to 800 μl of Z-buffer. To each of these tubes, 20 μl of 0.1% SDS and 50 μl of chloroform were added. The tubes were then vortexed for 15 seconds each to break open the cells, and incubated for 15 minutes at 30 °C in a water bath. To each of the tubes, 200 μl of 4 mg/ml of the pre-warmed ONPG was added and the tubes were vortexed for 5 seconds each, and placed back into the 30 °C water bath. The start time of the incubation was recorded. The tubes were observed for a visible colour change to yellow. Whenever colour change occurred, the time was recorded and 500 μl of 1M Na_2CO_3 was added to stop the reaction. If colour change was not immediate, the tubes were incubated at 30 °C overnight, until colour change was observed. After stopping the reaction, the tubes were centrifuged at 4000 RPM for 4 minutes, and the supernatant was transferred to a clean cuvette. The absorbance of each tube were measured at the wavelengths of 420 nm and 550 nm with a mixture of 1 ml Z-buffer and 0.5 ml 1M Na_2CO_3 as the blank. OD_{420} is the absorbance of the ONPG and cell debris, and OD_{550} is the absorbance of the cell debris alone.

3 Results

3.1 PIPE Results

A modified version of PIPE was developed to predict the precise sites of PPI in a high throughput fashion (Amos-Binks et al., 2011; Appendix A). This tool produced six different lists of motif pairs based on the selection criteria from the top 1000 predicted pairs of interactions. Much of each list was comprised of sequences with relatively low complexity containing large number of repeats. From the “1 novel” (one of the motifs is a novel motif) and “2 novel” (both motifs are novel) lists, a series of motifs were identified for further study. The total number of motifs for each category is listed in Table 2. The motifs selected for further study are listed in Table 5. These motifs contain most of our initially designed selection criteria. In addition complexity was used as an additional selection measure among the member of the same category. Both high and low complexity motifs were included in the final list.

Criteria	# of Motif Pairs
1 Novel	1138
1 Novel good length	119
1 Novel near end	232
1 Novel good length near end	35
2 Novel	271
2 Novel good length	20
2 Novel near end	80
2 Novel good length near end	6

Table 2: Total number of motifs generated by PIPE from the top 1000 PPI hits.

The 1 novel values includes the 1 novel good length, 1 novel near end, and 1 novel good length near end. The same also applies to the 2 novel categories.

The number of single amino acid repeats was also identified from the PIPE results. Repeats greater than five residues for each amino acid were totalled from both the one novel and two novel sets (Table 3). A minimum number of five was used based on previous studies that showed repeats as low as five were implicated in disease (Labaj et al., 2010). Repeats of aspartic acid, glutamic acid, asparagine, and glutamine were found in higher relative abundance, at 151, 151, 138, and 769 occurrences respectively. In addition, the abundance of these repeats in near-end motifs was examined, based on prior studies that showed that terminal repeats had functional importance (Coletta et al., 2010). The top 100 scoring near-end PIPE predicted pairs were compared with the total near-end motifs pairs (Table 4). A higher PIPE score indicates a greater frequency of motif co-occurrence in databases of protein pairs, indicating a greater representation of a particular amino acid repeat. Glutamine was found in the highest abundance, occurring 155 times in the top 100 near-end pairs compared to 1005 times in the total overall number of motif pairs.

	# of occurrences of repeats at least 5 amino acids in length from 1 novel sets	# of occurrences of repeats at least 5 amino acids in length from 2 novel sets
Arginine (R)	0	0
Histidine (H)	6	5
Lysine (K)	9	0
Aspartic Acid (D)	151	12
Glutamic Acid (E)	151	42
Serine (S)	30	18
Threonine (T)	0	0
Asparagine (N)	138	54
Glutamine (Q)	769	236
Cysteine (C)	0	0
Glycine (G)	0	0
Proline (P)	1	3
Alanine (A)	1	0
Valine (V)	0	0
Isoleucine (I)	0	0
Leucine (L)	0	0
Methionine (M)	0	0
Phenylalanine (F)	0	0
Tryptophan (Y)	0	0
Tyrosine (W)	0	0

Table 3: Representation of single amino acid repeats in PIPE results.

The table shows the number of single amino acid repeats greater than five in length that occurred in the top 1000 PIPE predicted interactions. Values do not represent unique motifs, but rather the frequency that motifs containing these sequences are represented in the top 1000 interactions. For example, sequences of five or greater glutamines (QQQQQ) occurred 769 times in the top 1000 interaction pairs.

	# of occurrences of repeats at least 5 residues in length, all 1409 motif pairs	# of occurrences of repeats at least 5 residues in length, all 312 near-end motif pairs	# of occurrences of repeats at least 5 residues in length, top 100 near-end motif pairs
Histidine (H)	11	3	0
Aspartic Acid (D)	163	11	2
Glutamic Acid (E)	193	36	9
Serine (S)	48	2	0
Asparagine (N)	192	83	11
Glutamine (Q)	1005	276	155
Proline (P)	4	2	0
Alanine (A)	1	0	0
Serine (S)	48	2	0
Lysine (K)	9	0	0

Table 4: Single amino acid repeats in top 100 PIPE pairs from near-end list of motifs.

The table shows the number of single amino acid repeats occurrences in PIPE predictions from the top 100 pairs of near-end motifs compared with all the near-end motifs and the total list of motif pairs. Values were obtained from aggregating all 1 novel and 2 novel near-end lists, which were sorted based on their PIPE score. A higher PIPE score indicates that the motif occurs more frequently in interacting proteins.

Yeast Protein Name	Motif Name	Amino Acid Position	Motif Sequence	Known Motif
YDL153C, Sas10p	m718 (novel)	60-90	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YNL154C, Yck2p	m719 (known)	182-207	IEDLHAHDLIYRDIKPDNFIIGRPG	Serine/Threonine protein kinases active-site (191-203)
YDL153C, Sas10p	m718 (novel)	60-90	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YDR247W, Vhs1p	m999 (known)	167-199	QICSALNYCHEHGIYHCDIKPENLLDTEENV	Serine/Threonine protein kinases active-site (181-193)
YDL153C, Sas10p	m718 (novel)	60-90	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YDR247W, Vhs1p	m1000 (known)	7-30	RINNYLITSQIGEGAYGLVYRAL	Protein kinases ATP-binding region signature (18-41)
YDL153C, Sas10p	m718 (novel)	60-90	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YER123W, Yck3p	m822 (known)	137-165	VRAIHDHDLIYRDIKPDNFIISQYQRIS	Serine/Threonine protein kinases active-site (146-158)
YDL153C, Sas10p	m1068 (novel)	60-91	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YER123W, Yck3p	m1069 (known)	11-33	IHYAVGPKIGEGSFGVIFEGEN	Casein kinase-related (2-468)
YDL153C, Sas10p	m1068 (novel)	60-91	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YER013W, Prp22p	m2181 (known)	347-368	LEEQRDETDEIDVELNTDDG	ATP-dependent RNA helicase (32-1141)
YDL153C, Sas10p	m1068 (novel)	60-91	LAMDEDDESIDEREDEEEEEEELDGA	(novel)
YER013W, Prp22p	m2183 (known)	500-523	FLVIVGETGSGKTTQITQYLDEE	Superfamilies 1 and 2 helicase ATP-binding type-I domain (493-656)
YHR030C, Slt2p	m108 (known)	132-170	FTYQILCGLKYIHSADVLHRDLKPGNLLVNADCQLKIC	Serine/Threonine protein kinases active-site (149-161)
YPL049C, Dig1p	m2164 (novel)	57-82	EHVPEEDDSGDKEADHEDSETATAK	(novel)
YER111C, Swi4p	m1955 (novel)	703-737	KLARNLPQKNYYQQQQQQQPQNNVKIPKIIKTQ	(novel)
YPR054W, Smk1	m637 (known)	145-178	FLYQILCGLKYIHSADVHRDLKPGNICTLNG	Serine/Threonine protein kinases active-site (162-174)
YER111C, Swi4p	m1956 (novel)	329-362	NNLIIVPDGPMQSQQQQHHHEYLNNFNHSM	(novel)
YPR054W, Smk1	m2187 (known)	145-188	FLYQILCGLKYIHSADVHRDLKPGNICTLNGCLKICDFGLA	Serine/Threonine protein kinases active-site (162-174)
YER111C, Swi4p	m1955 (novel)	703-737	KLARNLPQKNYYQQQQQQQPQNNVKIPKIIKTQ	(novel)
YPR054W, Smk1	m796 (known)	159-190	ADVIHRDLKPGNICTLNGCLKICDFGLARG	Serine/Threonine protein kinases active-site (162-174)
YGR086C, Pil1p	m1683 (novel)	272-298	SFKQDYEDFEPEEEEEEDGQGRW	(novel)
YLR113W, Hog1p	m536 (known)	122-160	YFLYQILRGLKYVHSAGVIHRDLKPSNILINENDLKI	Serine/Threonine protein kinases active-site (140-152)
YGR086C, Pil1p	m2188 (novel)	296-324	RWSEDEQEDGQIEEPEQEEEGAVEEHEQ	(novel)
YLR113W, Hog1p	m536 (known)	122-160	YFLYQILRGLKYVHSAGVIHRDLKPSNILINENDLKI	Serine/Threonine protein kinases active-site (140-152)
YGR086C, Pil1p	m430 (novel)	281-307	EPEEGEEEEEDGQGRWSEDEQEDGQ	(novel)
YLR113W, Hog1p	m2189 (known)	122-159	YFLYQILRGLKYVHSAGVIHRDLKPSNILINENDLK	Serine/Threonine protein kinases active-site (140-152)

Table 5: Motifs selected for further analysis using Y2H system.

The description of known motifs or domains in the PIPE predicted interaction pair is as described by Interpro queries. The amino acid positions of the predicted motifs were provided by PIPE, while the positions for the known motifs were from Interpro annotations. A 'known' motif constituted an overlap of the PIPE predicted motif and the Interpro annotated position. Common names of proteins were obtained from Saccharomyces Genome Database.

Motifs m718 and m1068 only differed by a single alanine on the C-terminal end, consequently they were assumed to be able to interact with the same set of partners. Protein names were obtained from SGD. The motifs that were found to be “known” (previously reported) by PIPE at the specified amino acid position were evaluated using Interpro and UniProt. The known domain descriptions were taken from Interpro annotations. The majority of the identified motifs belonged to protein kinases and the sites of interaction were found to be the kinase active sites or ATP-binding regions. A few of the known sequences did not belong to a specified motif or domain, but rather to a protein family. Sequences extracted from PIPE were filtered against Interpro to determine novelty. Interpro is considered a very inclusive database which integrates all protein sequence information from various sources, including one known as PANTHER. PANTHER, or **P**rotein **A**Nalysis **T**Hrough **E**volutionary **R**elationships, classifies genes and their products by their functions determined experimentally, and in some cases using evolutionary relationships (Dong et al., 2010). Because of this, it does not represent shorter domains, but instead the entire protein sequence of a protein.

3.2 PCR and Cloning for Bait and Prey Plasmids

The above motifs were generated using PCR with the protocol provided in Materials and Methods (Figure 3). Sequences of approximately 100-150 base pairs were visible compared to the control lanes, which only had primer dimers of roughly 60 base pairs.

Figure 3: PCR of motifs from S288c genomic DNA.

Lanes 1 and 2 are m718 (A). Lanes 4 and 5 are m1068 (B). Lanes 7 and 8 are m719 (C). Lanes 10 and 11 are m999 (D). Lanes 13 and 14 are m1000 (E). Lanes 16 and 17 are m822 (F). Lanes 19 and 20 are m1069 (G). Lanes 22 and 23 are m2181 (H). Lanes 25 and 26 are m2183 (I). Lanes 3, 6, 9, 12, 15, 18, 21, 24, and 27 are negative controls, containing no template DNA.

The motifs m718, m1068, m108, m1955, m1956, m1683, and m2188 were cloned into the pBI880 plasmid containing the binding domain, and m719, m999, m1000, m822, m1069, m2181, m2183, m2164, m637, m2187, m796, m536, m2189 were cloned into pBI881 containing the activation domain. Following purification, ligation, and transformation of ligated bait and prey vectors into DH5 α competent cells, colony PCR was performed to screen for correct insertion of the motif into pBI880 and pBI881 plasmids (Figure 4). Lanes 1 to 10 represent individual colonies picked, while lane 11s and 12s were the negative controls which had no cells added to the PCR reaction, and only had primer dimers forming. Clones that were suspected to be correct had approximately 100 base pair bands formed.

To confirm that colony PCR results were correct, and that the band produced wasn't part of the multiple cloning site from empty plasmids, up to six clones that showed a band for each motif were selected for screening using restriction digestion with NotI and Sall (Figure 5). Motif bands were relatively faint on the gel compared to the vector band, as they were much smaller and hence much lower in nucleotide content. For some of the motifs, such as C and E, the motif bands were very faint, but still visible.

Figure 4: Colony PCR for screening of correct motif into pBI880 and pBI881.

Digestions were run on a 1.5% agarose gel with a 100 base pair ladder. (A) m718, (B) m1068, (C) m719, (D) m999, (E) m1000, (F) m822, (G) m1069, (H) m2181. Lanes 1 through 10 were picked colonies while 11 and 12 were negative controls. Correct bands were identified at roughly 100 base pairs.

Figure 5: Restriction digests of plasmids containing inserted motifs.

Plasmids containing motif inserts were digested with NotI and SalI. Digestions were run on a 1.5% agarose gel with a 100 base pair ladder. (A) m718, (B) m1068, (C) m719, (D) m999, (E) m1000, (F) m822, (G) m1069, (H) m2181, (I) m2183. The larger bright bands were the vector, and very faint motif bands were visible at around 100 base pairs. The difference in visibility is due to the size disparity between the vector and motifs. Vector bands are roughly 80 times larger than the motifs. Negative controls used were empty pBI880 and pBI881 vectors digested with NotI and SalI.

|

3.3 Sequencing Results

Sequencing was performed to determine which of the pBI881 motif plasmids had been taken up by the cell during the second yeast transformation performed with a plasmid mixture. Sequences are listed in Appendix B for motif pairs that demonstrated positive interactions.

3.4 Yeast Two-hybrid Results

Yeast containing bait and prey plasmids with PIPE predicted motif pairs were spotted onto selection media. Figures 6 – 10 are spot test results on media lacking uracil. Colony growth in the absence of uracil is an indicator for the presence of a positive physical interaction between the two cloned motifs. In all cases, following incubation for two days there was significantly increased growth of colonies over the negative controls. Negative controls used were motifs in the pBI880 plasmid containing the binding domain alone. This was to ensure that there was no auto-activation by the binding domain and motif fusion. The results suggest that all the tested pairs appear to have some degree of interaction.

Figure 6: Spot test for motifs m1068/m719.

The empty plasmid pBI881 and motif m1068 was used for the negative control. Plates were incubated for two days. Growth of cells transformed with both bait and prey plasmids were observed after one day, and difference in size of colonies was more pronounced on day two.

Figure 7: Spot tests for motifs m1068/m822 and m1068/m2181.

Spot test for motifs m822 and m2181. The empty plasmid pBI881 and motif m1068 was used for the negative control. Plates were incubated for two days. Growth of cells transformed with both bait and prey plasmids were observed after one day, and difference in size of colonies was more pronounced on day two.

Figure 8: Spot test for motifs m1955/m637 and m108/m2164.

The empty plasmids pBI880 and pBI881 together were used for the negative control.

Plates were incubated for two days. Growth of cells transformed with both bait and prey plasmids were observed after one day, and difference in size of colonies was more pronounced on day two.

Figure 9: Spot test for motifs m1683/m536.

The empty plasmid pBI881 and motif m1683 was used for the negative control. Plates were incubated for two days. Growth of cells transformed with both bait and prey plasmids were observed after one day, and difference in size of colonies was more pronounced on day two.

Figure 10: Spot test for motifs m2188/m536.

The empty plasmid pBI881 and motif m2188 was used for the negative control. Plates were incubated for two days. Growth of cells transformed with both bait and prey plasmids were observed after one day, and difference in size of colonies was more pronounced on day two.

3.5 β -galactosidase Assay

Motif pairs that showed growth on –Ura selection media were measured for the amount of β -galactosidase production compared to the negative control. Production of β -galactosidase represents a positive interaction between the studied motif pairs. Empty plasmids were used as the negative control, which represented no interaction and a baseline level of enzyme production. Two genes previously cloned into bait and prey plasmids (TON3 and NPR1) that have been confirmed to strongly interact with each other were used as a positive control to ensure that the assay was performed correctly. The results of the assay are displayed in Figure 11. Three pairs of motifs did not show any significant increase over the negative control.

Figure 11: Relative β -galactosidase activity.

The amount of β -galactosidase activity was related to a negative control (pB1880 + pB1881) consisting of empty bait and prey plasmids, which demonstrated no interaction and change of absorbance, set at 1.00. TON3 and NPR1 are two proteins previously demonstrated to have strong interaction with each other, and were used to ensure the assay was functioning correctly. Each experiment was repeated three times or more with 15% or less variation and the average values are represented.

4 Discussion

4.1 Analysis of Interacting Motifs and Parent Proteins

Many of the motifs generated from the PIPE predictions were from novel interactions (never been reported in literature before). In certain cases the proteins themselves have not been well-characterized and hence required further investigation to ascertain whether interactions between the two proteins could be plausible. On the basis of the novel motifs found to interact in this study, Table 4 shows some of the possible novel interactions between certain proteins.

Yeast Protein Name	Motif Name	Motif Sequence
YDL153C, Sas10p	m718 (novel)	LAMDEDDESIDEREDEEEEEEELDGA
YNL154C, Yck2p	m719 (known)	IEDLHAHDLIYRDIKPDNFLIGRPG
YDL153C, Sas10p	m718 (novel)	LAMDEDDESIDEREDEEEEEEELDGA
YER123W, Yck3p	m822 (known)	VRAIHDHDLIYRDIKPDNFLISQYQRIS
YDL153C, Sas10p	m1068 (novel)	LAMDEDDESIDEREDEEEEEEELDGA
YER013W, Prp22p	m2181 (known)	LEEQQRDETDEIDVELNTDDG
YHR030C, Slt2p	m108 (known)	FTYQILCGLKYIHSADVLHRDLKPGNLLVNADCQLKIC
YPL049C, Dig1p	m2164 (novel)	EHVPEEDDSGDKEADHEDSETATAK
YER111C, Swi4p	m1955 (novel)	KLARNLPQKNYYQQQQQQQPQNNVKIPKIIKTQ
YPR054W, Smk1p	m637 (known)	FLYQILCGLKYIHSADVIHRDLKPGNICTLNG
YGR086C, Pil1p	m1683 (novel)	SFKQDYEDFEPEEGEEEEEDGQGRW
YLR113W, Hog1p	m536 (known)	YFLYQILRGLKYVHSAGVIHRDLKPSNILINENCDLKI
YGR086C, Pil1p	m2188 (novel)	RWSEDEQEDGQIEEPEQEEEGAVEEHEQ
YLR113W, Hog1p	m536 (known)	YFLYQILRGLKYVHSAGVIHRDLKPSNILINENCDLKI

Table 6: Pairs of motifs that are shown to interact based on Y2H analysis.

The table shows the pairs of motifs that were shown to interact with each other using Y2H analysis as well as the parent yeast proteins that contain the motifs. The protein pairs represent interactions predicted by PIPE. Follow-up was conducted to determine whether these interactions had been previously been experimentally verified.

Sas10p, which stands for something about silencing protein, was originally identified as a novel protein that when overexpressed led to reduced transcriptional silencing in yeast strains (Kamakaka and Rine, 1998). Sas10p was later found to be part of the nuclear assembly protein family and an essential subunit of the U3-containing small subunit processome, which is required for the processing of the 35S pre-ribosomal RNA transcript into 18S ribosomal RNA (Dragon et al., 2002). It was also found to be involved in aneuploidy tolerance as a single amino acid mutation conferred increased proliferation in aneuploidy cells (Torres et al., 2010). This protein was found to interact with three different partners: Yck2p, Yck3p, and Prp22p. Yck2p and Yck3p are both casein kinases, and Yck2p has been found to be involved in septin assembly and endocytotic trafficking, while Yck3p has been implicated in vesicle fusion involved with the AP-3 transport pathway (Anand et al., 2009; Robinson et al., 1999). Our data suggest that both Yck3 and Yck2 compete to interact with the same site on Sas10p (Figure 12B). Prp22p is a DExH-box RNA helicase and is involved in pre-mRNA splicing (Schwer and Gross, 1998). While no literature information is available to indicate interactions between Sas10p and Yck3p or Prp22p, there has been a single study conducted that showed Yck2p is capable of phosphorylating Sas10p (Ptacek et al. 2005). This was from a large-scale study of protein phosphorylation in yeast using microarrays, where it was shown that radiolabeled ATP incorporation into a chip followed incubation with Yck2p. Yck3p has been shown to share overlapping functions with Yck2p and another casein kinase, Hrr25p (Wang et al., 1996). Evidence from genetic studies has shown that *YCK3* and *HRR25* form a gene pair with overlapping essential functions; *hrr25Δ yck3Δ* double mutant cells were found to be inviable, indicating that there is some overlap between

their two functions (Wang et al., 1996). Based on a large scale study to identify protein complexes in yeast using mass spectrometry, Hrr2p has also been shown to be able to interact with Sas10p (Ho et al., 2002). FLAG-tagged fusion bait proteins were overexpressed and purified using anti-FLAG resin, then subjected to high-throughput mass spectrometric protein complex identification, allowing for detection of any interacting proteins co-purified with the bait protein. Based on the sequence and functional similarities between Hrr2p and Yck3p, it may be possible that Sas10p is also an interacting partner with Yck3p (Figure 12A).

A.

Yck2p	HNILVIDLLGPSLEDLFDGCGRRFSVKTWVQVAVOMITLIEDLHAHDLIYRDIKPDNFLIGR-----PG	207
Yck1p	HNILVIDLLGPSLEDLFDGCGRRFSVKTWVQVAVOMITLIEDLHAHDLIYRDIKPDNFLIGR-----PG	200
Yck3p	HNILVIDLLGPSLEDLFDGCGRRFSVKTTCMVAKOMIDFVRATHDHLIYRDIKPDNFLISQYQRISPEGKVIKSCASS	178
Hrr25p	YNAMVIDLLGPSLEDLFDNYCHRRFSFKTVMIALOMFCRIQYIHGRSFIHRDIKPDNFLMCGVGR-----	141

B.

Figure 12: Binding model of Yck2p and Yck3p to Sas10p.

(A) The large box indicates PIPE predicted sequence on Yck3p to interact with Sas10p. Yck3p shares some homology with Yck1p and Yck2p, and to a lesser extent Hrr2p. Kinase active sites are position 146-158 for Yck3p, 191-203 for Yck2p, and 125-136 for Hrr25p, represented by the smaller boxes. Sas10p has been reported to interact with Yck2p and Hrr25p, but there has not been any prior evidence demonstrating interaction with Yck3p (Adapted from Wang et al., 1996). (B) Yck2p, Yck3p, and Hrr25p are suggested to interact with Sas10p at the same binding site in a competitive fashion. The three proteins have been previously shown to have overlapping function, and are able to compensate for each other in deletion strains.

Slt2p, a mitogen activated protein (MAP) kinase was found to interact with Dig1p, a MAP kinase responsive inhibitor. Dig1p is part of the invasive and filamentous growth pathway in yeast and regulation of mating-specific genes, and has been shown to inhibit the Ste12p transcription factor. Based on genetic studies, the loss of *DIG1* and its partner *DIG2* results in invasive growth by haploid strains that are normally unable to undergo this form of development (Cook et al., 1996). Dig1p is phosphorylated by Kss1p, another MAP kinase that is also part of pathways that control filamentous growth and pheromone response. Slt2p, also known as Byc2p, Mpk1p, and Slk2p, is a MAP kinase that is involved in several different pathways, including maintenance of cell wall and cell cycle progression (Madden et al., 1997; Martin-Yken et al., 2003). From affinity capture-immunoblot followed by mass spectroscopy, Slt2p has been shown to be able to directly interact with Dig1p (Breitkreutz et al., 2010). Our analysis of the sequence indicates that Dig1p interacts with Slt2p at a region contained within the substrate binding site of the kinase. While both of these proteins are from distinctly different pathways that seemingly have no connection, it has been shown in large-scale studies that Slt2p is able to interact with other targets that are part of the invasive growth pathway. Aside from Dig1p, Slt2p is able to interact with Ste12p, also an interactor with Dig1p, as well as Msn1p, Ste20p, Rxt2p, and Flo8p, which are all proteins that are suspected to be involved in invasive growth or pheromone response (Fiedler et al., 2009; Sharifpoor et al., 2012; Wojda et al., 2007). It is possible that Slt2p's kinase functions target many different substrates from several pathways (Figure 13).

Figure 13: Hypothetical interaction model for Slt2p.

Slt2p has been shown to interact with Ste12p, Ste20p, Msn1p, Rxt2p, and Flo8p in previous studies, all proteins implicated in invasive growth. PIPE predicts that Dig1p, also involved in the invasive growth pathway, is able to interact with Slt2p, indicated by the thicker line.

Swi4p is a protein that forms the DNA binding component part of the SBF complex, a transcription factor that is responsible for controlling the expression of key cell cycle regulatory proteins (Andrews and Moore, 1992). Its interacting partner, Smk1p, is a MAP kinase involved in outer spore wall assembly (Krisak et al., 1994). The two have not been previously shown to be able to interact with each other, and are from two completely separate pathways. However, Smk1p has been shown to physically interact with a variety of proteins from different pathways, including Cak1p (cell cycle), Mae1p (sugar metabolism), and Bud7p (cellular transport). Swi4p interacts with Smk1p's kinase active site, so it is possible that Swi4p represents a novel substrate.

The final sets of motifs are found in Hog1p and Pil1p. Hog1p is a MAP kinase involved in the osmoregulation signalling pathway, which affects gene expression and cell cycle in response to hyperosmotic stress (Brewster et al., 1993). It is the yeast homolog of the mammalian MAP kinase p38, which is an important part of inflammatory and stress responses (Han et al., 1993). Pil1p is a key component of eisosomes, which are structures responsible for mediating endocytosis in yeast cells (Walther et al., 2006). Originally believed to be a random process, it was discovered that endocytosis forms at these eisosomes, which are large immobile protein complexes formed from Pil1p and Lsp1p. There is no prior evidence of Hog1p interacting with Pil1p. However, Pil1p has been shown to be phosphorylated by other serine/threonine kinases, including Pkh1p and Pkh2p. When we compare our motif sequence from Hog1p to Pkh1p and Pkh2p, we find a high degree of similarity, which was not surprising as they all share roles as serine/threonine kinases and should share conserved regions (Figure 14). The mechanisms of how Pil1p interacts with Pkh1p and Pkh2p are unknown. However, based

on our findings it may be possible that Pil1p interacts with Pkh1p and Pkh2p at the sequences homologous to our PIPE-predicted motif on Hog1p. The two theorized binding sequences from Pil1p are novel, and could represent necessary motifs required for phosphorylation by other serine/threonine kinases.

A.

Hog1p	116	--LEKQFVQYFLYQILRGLKYVHSAGVIHRDLKPSNILINENCDLKI	CDF	163
		: :		
Pkh1p	219	GSLNETCARYYASQIIDAVDSLHNIGIIHRDIKPENILLDKNMKVKI	TDF	268

B.

Hog1p	113	TRP-----LEKQFVQYFLYQILRGLKYVHSAGVIHRDEKPSN	149	
		.. :		
Pkh2p	259	YAPNGDFLSLMKKYGSLETCARYYAAQIIDAYLHSGIIHRDIKPEN	308	
Hog1p	150	ILINENCDLKI	CDFGLARIQDP-----QMTGYVSTRYYRAP	185
		: :		
Pkh2p	309	ILLDGEKIKL	TDFGTAKLLNPTNNSVSKPEYDLSTRSKSFVGTAEYVSP	358

Figure 14: Sequence alignments between Hog1p and Pkh1p/Pkh2p.

The protein sequences for Hog1 were aligned pairwise with (A) Pkh1p and (B) Pkh2p. The highlighted regions represent our PIPE-predicted motif and the homologous sequences on Pkh1p and Pkh2p. For Hog1p and Pkh1p there is 73.7% similarity and 42.1% identity between the highlighted sequences and 71.1% similarity and 44.7% identity for Hog1p and Pkh2p.

When looking at the proposed interactions between Slt2p and Swi4p and Hog1p and Pil1p there doesn't seem to be any evidence linking the two interacting proteins together. However, it must be considered that signalling pathways do not function independently of one another and need to communicate with each other. Hog1p for example, is an important component of osmotic stress response. The full extent to which cells responds to changes in osmolality are unknown, but it has been suggested that an increase in solute concentration changes the behaviour of plasma membrane dynamics, including endocytosis. In experiments conducted where cells were submersed in high levels of glycerol to induce hyperosmolarity, arrest of endocytosis was observed (Dupont et al., 2010). While the mechanisms for how this occurs are unknown, the pathways involved in osmotic regulation and endocytosis needs to be able to communicate to facilitate this type of coordinated response. The novel interaction between Hog1p and Pil1p that we predicted in this study may serve as a communication bridge between these two seemingly independent pathways. It is also suggested that the common theme among all yeast MAP kinases is an involvement in the control of cell growth and morphogenesis in response to different cellular stimuli (Hohmann, 2010). If this is the case, then it is possible that yeast MAP kinases are able to directly interact with many different pathways and there would be significant cross communication between those pathways.

4.2 Amino Acid Sequence Complexity

Many of the motifs generated by PIPE consisted of low complexity sequences (LCS) comprised of single amino acid repeats. When run through BLAST with the SEG low complexity filter turned on, these sequences were discarded by the filter. For the

purpose of this study, these sequences were initially disregarded as simply random peptide sequences. This is due to the long standing assumption that these low complexity sequences are the equivalent of “junk” DNA and have poorly understood or lack any functional roles. Sequence complexity is representative of the traditional paradigm of linking three-dimensional structure to biological function, and any non-ordered sequence is the result of neutral or random evolution. However, there has been increasing study into trying to define a role for these simple sequences and single amino acid repeats. These simple sequences have been found in large abundance in a variety of organisms ranging from roughly 15% of all *S. cerevisiae* and *H. sapiens* genes containing at least one low complexity sequence, to over 50% in *Dictyostelium discoideum*, a species of soil eating amoeba (Eichinger et al., 2001).

It is observed that low complexity sequences consist primarily of polar and acidic amino acids, with single amino acid amino repeats composed of glutamine, asparagine, serine, glycine, and alanine (Haerty and Golding, 2010). Repeats of isoleucine, methionine, tryptophan, tyrosine, and cysteine are rare. This is consistent with the PIPE results, as out of the 1139 single novel and the 272 double novel interaction pairs, repeats of glutamine and asparagine were found to be among the most common. Repeats of aspartic acid and glutamic acid were also found to be very common, which is consistent with previous findings that repeats of codons for these amino acids are found in relative higher frequency in yeast. Out of the total *S. cerevisiae* genome, the codons for aspartic acid and glutamic acid were repeated at least 7 times in 81 occurrences for each (Katti et al., 2001).

It has been found that proteins with a large number of low complexity regions, or LCRs, have a greater number of interacting partners compared to those without, and that these sequences tend to be biased towards the outer ends of protein amino acid sequence (Coletta et al., 2010). The same study, using gene ontology enrichment analysis with the *S. cerevisiae* proteome, also found that proteins with low complexity regions tended be enriched for functions related to transcription, transcription regulation, and translation. Furthermore, it was also found that LCRs located towards the ends of proteins had differing functions compared to those found toward the centre of protein sequences. Terminal LCRs, or t-LCRs, were defined as any region that started or ended no more than 25 amino acids from the protein termini, and proteins containing t-LCRs were found to be important to stress response, translation, and transport processes. Central LCRs, or c-LCRs, were considered to be regions that were located at least 50 amino acids from the protein sequence ends. We found c-LCRs were more likely involved with transcription and transcription regulation processes. LCRs were also enriched for kinase function. It was hypothesized that these end terminal low complexity regions are involved with low-specificity interaction events that involve large protein complexes. This was from the basis that disordered lengthy single amino acid repeats can extend free from the rest of the protein structure, thus acting as a flexible and promiscuous interface for protein binding. We analyzed our PIPE data to determine whether there was an increased occurrence of low complexity sequences at the terminal ends (Table 4). Based on our results, only repeats of glutamine appeared to be enriched in near-end motifs. This may be significant based on evidence suggesting that glutamine repeats in particular have an important role in protein-protein interactions. The expansion of glutamine repeats has

long been observed in the genes of many patients suffering from genetic diseases such as Huntington's disease, a disease characterized by excessive protein aggregation (Gatchel and Zoghbi, 2005). Repeats of glutamine residues are thought to be able to stabilize PPIs, and increased length of the repeats has been shown to result in abnormal interactions (Schaefer et al., 2012). It was also discovered that proteins with glutamine repeats have on average higher numbers of interacting partners compared to other proteins, which is consistent with our PIPE results and would explain why the motif pairs with higher scores had relatively higher amounts of glutamine repeats. It has been proposed that these repeats form a motif for protein-protein interactions. While there has been much study into glutamine repeats and their effect on protein aggregation in diseases, the role that these glutamine repeats have in mediating PPIs in healthy cells may warrant further investigation.

4.3 Drug Discovery and Protein-Protein Interactions

Knowing the exact binding sites of PPIs is crucial for targeted drug design. Because PPIs permeate all facets of cellular mechanisms, it is understandable that many diseases are caused by improper interaction states. Diseases such as Alzheimer's have been linked to excessive aggregation of certain proteins and a number of cancers are due to increased activity of key kinases. Over the past several years there has been a large amount of effort in designing novel drugs that can disrupt protein interactions by targeting the binding sites with high specificity and affinity. Design of these drugs requires a starting structure that is known and can be modified or replicated synthetically, either by using a small peptide or peptidomimetic (Arkin and Wells, 2004). To a certain

extent this has been successful in some instances, such as disruption of human HER2-EGFR and HER2-HER3 interactions that are implicated in forms of breast cancer using a compound that binds to the domain IV of HER2 (Banappagari et al., 2012). This was partly possible due to availability of a detailed molecular structure of HER2. For most proteins, this isn't an option and using predictive software such as PIPE to determine possible starting sequences may be useful.

Discovery of PPI inhibitors poses several challenges. Perhaps the primary issue is that many interaction sites between globular proteins are large and relatively flat, with a number of weak interaction points that sum up to create a stronger interaction (Fry, 2006). A single small molecule would have difficulty duplicating every one of those interactions. However, it has been found that many interactions contain a 'hotspot'—a subset of residues that consists of less than half of the binding interface of the interaction and contribute the majority of the free energy necessary for binding. These hotspots are also usually found in the centre of the contact surface. Identification of these critical residues can provide insight into the optimization of corresponding ligands (Moreira et al., 2007). PIPE provides a good starting point for determining these critical residues. From our investigation it seems possible that short motifs can alone facilitate binding, which would be in line with the idea that a few important residues can provide the necessary energy to maintain interaction. Even if PIPE predicts a highly unlikely interaction *in vivo* between two proteins from two very different pathways, follow-up studies that demonstrate some level of interaction are still useful. This indicates that the addition of the PIPE predicted sequences to the binding and activating domains adds residues that can mediate interaction. From this starting point, it is possible to use

mutagenesis studies to find the exact critical residues. As well, PIPE often predicted multiple sequences per protein pair that were responsible for interactions. Based on what is known about interaction interface, this is not surprising, as many contact sites are not made up of single contiguous sequences. Instead, multiple points on a sequence fold up to form a 2D binding surface.

There is growing evidence to suggest that many interactions between globular proteins can be mediated by short linear motifs that bind to a specific region on a protein or domain (Neduva and Russell, 2006). Serine/threonine kinases in particular are regulated by these types of interactions and discovery of kinase inhibitors is an active area of research, especially the search for new docking peptide sequences (Eldar-Finkelman and Eisenstein, 2009). Kinases share a common catalytic fold, but have incredible substrate binding diversity due to other sequences or regulatory domains. In our investigation, several of the motif pairs belonged to a serine/threonine kinase substrate interaction. For example, motif m822 (VRAIHDHDLIYRDIKPDNFLISQYQRIS) belongs to the active site of kinase Yck3p, and was found to be able to interact with motif m1068 (LAMDEDDESIDEREDEEEEEEELDGA). PIPE also predicted six other motifs that could interact with m822. It is possible that one of those motifs binds to Yck3p with high affinity, and could serve as a template for designing a competitive inhibitor, which would prevent the substrate docking. To determine whether these interactions could change kinase activity, further activity analysis would be needed. However, it is important to consider that binding specificity is crucial for designing targeted drugs. Motif m718 was predicted to interact with three different serine/threonine kinases at five different binding

sites which were all confirmed kinase domains. While not all of them displayed high affinity interactions during our Y2H analysis, this could have been due to our experimental design rather than an actual false interaction. Because only the PIPE predicted sequences were cloned into two-hybrid plasmids, a fully functional kinase domain binding pocket may not have been completely produced, which would not result in sufficient space for interaction with motif m718. Further analysis should be carried out using the full kinase domains or full length globular protein to provide an accurate representation of the binding conditions within the cell. In this instance, more research would have to go into finding the residues responsible for mediating interactions, and determining whether changing the sequence can alter specificity.

4.4 Conclusion and Future Directions

Using PIPE we were able to confirm several novel motif-based PPI predictions while discovering a handful of possible new motifs that can mediate interactions. For this study we chose to use Y2H as the method for confirming predicted interaction between motifs, and while we were successful, it still only represents a minute fraction of the data that PIPE is able to produce. Many other factors are still worth examining. The number of low complexity sequences and their roles in mediating interactions is worth investigating further. It may also be useful to determine the exact residues responsible for PPIs using mutagenesis studies, as well as how these changes can affect enzymatic activity. As well, PIPE often predicted multiple sequences on the same protein pair. While our analysis was based on the idea that single sequences could mediate interactions, it is worth noting that interactions may be more complex, involving multiple

sites on a single protein. In our case we were able to see interaction using pairs of single motifs; however in reality for true interaction to occur multiple motifs may need to be present.

The more we learn about PPIs and their role in cellular function and disease, the greater the apparent complexity of interaction networks becomes. PIPE is a powerful tool for studying PPIs and one of the many new approaches that exists for tackling the challenging task of studying the interactome.

Appendices

Appendix A Published Paper from Thesis

Amos-Binks et al. *BMC Bioinformatics* 2011, 12:225
<http://www.biomedcentral.com/1471-2105/12/225>

METHODOLOGY ARTICLE

Open Access

Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences

Adam Amos-Binks^{1†}, Catalin Patulea^{3†}, Sylvain Pitre¹, Andrew Schoenrock¹, Yuan Gui², James R Green³, Ashkan Golshani² and Frank Dehne^{1*}

Abstract

Background: While there are many methods for predicting protein-protein interaction, very few can determine the specific site of interaction on each protein. Characterization of the specific sequence regions mediating interaction (binding sites) is crucial for an understanding of cellular pathways. Experimental methods often report false binding sites due to experimental limitations, while computational methods tend to require data which is not available at the proteome-scale. Here we present PIPE-Sites, a novel method of protein specific binding site prediction based on pairs of re-occurring polypeptide sequences, which have been previously shown to accurately predict protein-protein interactions. PIPE-Sites operates at high specificity and requires only the sequences of query proteins and a database of known binary interactions with no binding site data, making it applicable to binding site prediction at the proteome-scale.

Results: PIPE-Sites was evaluated using a dataset of 265 yeast and 423 human interacting proteins pairs with experimentally-determined binding sites. We found that PIPE-Sites predictions were closer to the confirmed binding site than those of two existing binding site prediction methods based on domain-domain interactions, when applied to the same dataset. Finally, we applied PIPE-Sites to two datasets of 2347 yeast and 14,438 human novel interacting protein pairs predicted to interact with high confidence. An analysis of the predicted interaction sites revealed a number of protein subsequences which are highly re-occurring in binding sites and which may represent novel binding motifs.

Conclusions: PIPE-Sites is an accurate method for predicting protein binding sites and is applicable to the proteome-scale. Thus, PIPE-Sites could be useful for exhaustive analysis of protein binding patterns in whole proteomes as well as discovery of novel binding motifs. PIPE-Sites is available online at <http://pipe-sites.cgmlab.org/>.

Background

In recent years, large-scale protein sequencing has generated a tremendous amount of data at the proteome level. However, biomedical significance of these results relies on the determination of protein function at this scale. Protein-protein interactions (PPIs) are the main mechanism behind biological pathways [1] and as such are at the core of a functional understanding of the cell under normal and pathological conditions. Characterization of the

sequence regions that give rise to protein binding sites responsible for interaction (interaction or binding site) opens the door to modulation of these pathways and is crucial to targeted drug design.

Large-scale protein-protein interaction (PPI) studies are bottlenecked by existing experimental and computational methods. Experimental assays such as yeast two-hybrid (Y2H) [2,3] or TAP-tagging [4,5] have been adapted to the proteome scale but lag in terms of throughput and robustness. For example, Y2H underreports interactions involving membrane and cytosol proteins because the assay requires the proteins to interact with the nucleus [6]. Deletion experiments, for binding site determination, are even more time-consuming and

* Correspondence: frank@dehne.net

† Contributed equally

¹School of Computer Science, Carleton University, Ottawa, ON K1S5B6, Canada

Full list of author information is available at the end of the article

© 2011 Amos-Binks et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Appendix B Sequencing Results

B.1 Motif m2181

TTAGTCTGTCACCGTTTCTTAGACCACTACATGGATGATGTATATAACTATCT
ATTCGATGATGAAGATACCCACCAAACCCAAAAAAGAGGGTGGGTCGAG
GGACTACAAGGACGACGACGACAAGGGGTTCGACTTGGGAAGAACAACAACGA
GACGAGACGGATGAAATTGATGTCGAGCTGAATACAGATGATGGGGCGGCC
GCTAAGTAAGTAAGACGTCGAGCTCTAAGTAAGTAACGGCCGCCACCGCGGT
GGAGCTTTGGACTTCTTCGCCAGAGGTTTGGTCAAGTCTCCAATCAAGGTTGT
CGGCTTGTCTACCTTGCCAGAAATTTACGAAAAGATGGAAAAGGGTCAAATC
GTTGGTAGATACGTTGTTGACACTTCTAAATAAGCGAATTTCTTATGATTTAT
GATTTTATTATTAAATAAGTTATAAAAATTATAAGTGTATACAAATTTTAAA
GTGACTCTTAGGTTTTAAACGAAACTCCTTATTCTTGAGTAATTCTT

B.2 Motif m822

GGCGTCATTTGCTGTTTCCTCGTGGATATAACCACTACAATGGATGATGTATAT
AACTATCTATTCGATGATGAAGATACCCACCAAACCCAAAAAAGAGGGTG
GGTCGAGGGACTACAAGGACGACGACGACAAGGGGTGACGTTAGAGCAAT
TCATGATCACGACTTAATCTATCGCGATATTAAACCCGATAACTTTTTAATTT
CTCAATATCAAAGAATTTCAAGCGGCCGCTAAGTAAGTAAGACGTCGAGCTCT
AAGTAAGTAACGGCCGCCACCGCGGTGGAGCTTTGGACTTCTTCGCCAGAGG
TTTGGTCAAGTCTCCAATCAAGGTTGTCGGCTTGTCTACCTTGCCAGAAATTT
ACGAAAAGATGGAAAAGGGTCAAATCGTTGGTAGATACGTTGTTGACACTTC
TAAATAAGCGAATTTCTTATGATTTATGATTTTTATTATTAAATAAGTTATAA
AAAAAATAAGTGTATACAAATTTTAAAGTGACTCTTAGGTTTTAAACGAAA
ATTCTTATTCTTGAGTAACTCTTTCCTGTAGGTCAGGTTGCTTCTCAGGTATA
GCATGAGGTCGCTCTTATTGACCACACCTCTACCGGCATGCCGAGCAAATGC
CTGCAAATCGCTCCCCATTTACCCAATTGTAGATATGCTAACTCCAGCAATG
AGTTGATGAATCTCGGTGTGTATTTTATGTCCTCAGAGGACAACACCTGTTGT
AATCGTTCTTCCACACGGATCCCAATTCGCCCTATAGTGAGTCGTATTACAAT
TCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCC
AACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCGCTGGCGTATTAGCGAAA
GGCCGCGCCGATCGCCCTTCCCACAGTGCCAGCCTGAATGGCGATGGACGCA
CCTGTAACGT

B.3 Motif m719

AATCGTGTACCAATGTTGTTTTATACCACTACAATGGATGATGTATATAACTA
TCTATTCGATGATGAAGATACCCACCAAACCCAAAAAAGAGGGTGGGTCG
AGGGACTACAAGGACGACGACGACAAGGGGTCGACATTGAAGATTTCACG
CTCATGATTTGATATACCGTGATATTAAACCTGATAATTTTTTGATCGGAAGA
CCGGGTGCGGCCGCTAAGTAAGTAAGACGTCGAGCTCTAAGTAAGTAACGGC
CGCCACCGCGGTGGAGCTTTGGACTTCTTCGCCAGAGGTTTGGTCAAGTCTCC
AATCAAGGTTGTCGGCTTGTCTACCTTGCCAGAAATTTACGAAAAGATGGAA
AAGGGTCAAATCGTTGGTAGATACGTTGTTGACACTTCTAAATAAGCGAATTT
CTTATGATTTATGATTTTTATTATTAAATAAGTTATAAAAAAATAAGTGTAT
ACAAATTTT

References

- Alberghina, L., & Cirulli, C. (2010). Proteomics and systems biology to tackle biological complexity: Yeast as a case study. *Proteomics*, 10(24), 4337-41.
- Aloy, P., & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), 5896-901.
- Amos-Binks, A., Patulea, C., Pitre, S., Schoenrock, A., Gui, Y., Green, J. R., Golshani, A., et al. (2011). Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC Bioinformatics*, 12(1), 225.
- Anand, V. C., Daboussi, L., Lorenz, T. C., & Payne, G. S. (2009). Genome-wide Analysis of AP-3–dependent Protein Transport in Yeast. *Molecular Biology of the Cell*, 20(5), 1592-1604.
- Andrews, B. J., & Moore, L. a. (1992). Interaction of the yeast Swi4 and Swi6 cell cycle regulatory proteins in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 89(24), 11852-6.
- Andrews, B., & Measday, V. (1998, February). The cyclin family of budding yeast: abundant use of a good idea. *Trends in Genetics : TIG*. 14(2), 66-72

- Arkin, M. R., & Wells, J. a. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Reviews. Drug Discovery*, 3(4), 301-17.
- Banappagari, S., Corti, M., Pincus, S., & Satyanarayanajois, S. (2012). Inhibition of protein-protein interaction of HER2-EGFR and HER2-HER3 by a rationally designed peptidomimetic. *Journal of Biomolecular Structure & Dynamics*, (July 2012), 37-41.
- Bennett, M., & Schlunegger, M. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Science*, 4(12), 2455-2468.
- Bennetzen, J. L., & Hall, B. D. (1982). Codon selection in yeast. *Journal of Biological Chemistry*, 257(6), 3026-3031.
- Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A., & Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences: CMLS*, 62(4), 435-45.
- Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.-Y., et al. (2010). A global protein kinase and phosphatase interaction network in yeast. *Science (New York, N.Y.)*, 328(5981), 1043-6.
- Brewster, J. L., de Valoir, T., Dwyer, N. D., Winter, E., & Gustin, M. C. (1993). An osmosensing signal transduction pathway in yeast. *Science (New York, N.Y.)*, 259(5102), 1760-3.

- Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10(6), 2763-88.
- Bähler, J. (2005). Cell-cycle control of gene expression in budding and fission yeast. *Annual Review of Genetics*, 39, 69-94.
- Causier, B., & Davies, B. (2002). Analysing protein-protein interactions with the yeast two-hybrid system. *Plant Molecular Biology*, 50(6), 855-70.
- Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L., & Cesareni, G. (2007). DOMINO: a database of domain-peptide interactions. *Nucleic Acids Research*, 35(Database issue), D557-60.
- Chevray, P. M., & Nathans, D. (1992). Protein interaction cloning in yeast: identification of mammalian proteins that react with the leucine zipper of Jun. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13), 5789-93.
- Chothia, C., & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256(8), 705-708.
- Clackson, T. and Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196), 383-386.

- Coletta, A., Pinney, J. W., Solís, D. Y. W., Marsh, J., Pettifer, S. R., & Attwood, T. K. (2010). Low-complexity regions within protein sequences have position-dependent roles. *BMC Systems Biology*, 4(43), 1-13.
- Cook, J. G., Bardwell, L., Kron, S. J., & Thorner, J. (1996). Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast *Saccharomyces cerevisiae*. *Genes & Development*, 10(22), 2831-2848.
- Copley, R. R., Doerks, T., Letunic, I., & Bork, P. (2002). Protein domain analysis in the era of complete genomes. *FEBS Letters*, 513(1), 129-34.
- Cusick, M. E., Klitgord, N., Vidal, M., & Hill, D. E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(2), R171-81.
- Dragon, F., Gallagher, J. E. G., Compagnone-Post, P. A., Mitchell, B. M., Porwancher, K. A., Wehner, K. A., Wormsley, S., et al. (2002). A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature*, 417(6892), 967-70.
- Dupont, S., Beney, L., Ritt, J.-F., Lherminier, J., & Gervais, P. (2010). Lateral reorganization of plasma membrane is involved in the yeast resistance to severe dehydration. *Biochimica et Biophysica Acta*, 1798(5), 975-85.

- Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., et al. (2005). The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, 435(7038), 43-57.
- Eldar-Finkelman, H., & Eisenstein, M. (2009). Peptide inhibitors targeting protein kinases. *Current Pharmaceutical Design*, 15(21), 2463-70.
- Evans, J., & Manjunatha, R. (1996). Prediction of potential protein-protein interaction amino acid sequence Identification of a fibrin polymerization sites from site. *FEBS Letters*, 385, 81-86.
- Fiedler, D., Braberg, H., Mehta, M., & Chechik, G. (2009). Functional organization of the *S. cerevisiae* phosphorylation network. *Cell*, 136(5), 952-963.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., et al. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue), D281-8.
- Fry, D. (2006). Protein-protein interactions as targets for small molecule drug discovery. *Peptide Science*, 84(6), 535-552.
- Gatchel, J. R., & Zoghbi, H. Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics*, 6(10), 743-55.

- Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), 141-7.
- Ghose, R., Shekhtman, A., Goger, M. J., Ji, H., & Cowburn, D. (2001). A novel, specific interaction involving the Csk SH3 domain and its natural ligand. *Nature Structural Biology*, 8(11), 998-1004.
- Haerty, W., & Golding, G. B. (2010). Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. *Genome*, 53(10), 753-762.
- Han, J., Lee, J. D., Bibbs, L., & Ulevitch, R. J. (1994). A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells. *Science (New York, N.Y.)*, 265(5173), 808-11.
- Hegyí, H., & Gerstein, M. (2001). Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins. *Genome Research*, 11(10), 1632-1640.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G., & Moore, L. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(January), 180-183.
- Hooda, Y., & Kim, P. M. (2012). Computational structural analysis of protein interactions and networks. *Proteomics*, 12(10), 1697–1705.

- Hohmann, S. (2009). Control of high osmolarity signalling in the yeast *Saccharomyces cerevisiae*. *FEBS Letters*, 583(24), 4025-9.
- Huang, D., Friesen, H., & Andrews, B. (2007). Pho85, a multifunctional cyclin-dependent protein kinase in budding yeast. *Molecular Microbiology*, 66(2), 303-14.
- Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(Database issue), D306-12.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), 4569-74.
- Iyer, K., Burkle, L., Auerbach, D., & Thaminy, S. (2005). Utilizing the split-ubiquitin membrane yeast two-hybrid system to identify protein-protein interactions of integral membrane proteins. *Science STKE*, 2005(275), 1-25.
- Janin, J., Miller, S., & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology*, 204(1), 155-64.
- Jones, S., & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, 272(1), 121-32.

Jones, Susan, & Thornton, J. M. (1996). Principles of protein-protein interactions.

Proceedings of the National Academy of Sciences of the United States of America,
93(January), 13-20.

Kamakaka, R. T., & Rine, J. (1998). Sir-and silencer-independent disruption of silencing
in *Saccharomyces* by Sas10p. *Genetics*, 149(2), 903–914.

Kann, M. G. (2007). Protein interactions and disease: computational approaches to
uncover the etiology of diseases. *Briefings in Bioinformatics*, 8(5), 333-46.

Katti, M. V., Ranjekar, P. K., & Gupta, V. S. (2001). Differential distribution of simple
sequence repeats in eukaryotic genome sequences. *Molecular Biology and
Evolution*, 18(7), 1161-7.

Kini, R. M., & Evans, H. J. (1995). A hypothetical structural role for proline residues in
the flanking segments of protein-protein interaction sites. *Biochemical and
Biophysical Research Communications*, 212(3), 1115-1124.

Koike, A., & Takagi, T. (2004). Prediction of protein-protein interaction sites using
support vector machines. *Protein Engineering, Design & Selection : PEDS*, 17(2),
165-73.

Kohalmi, E. S., Nowak, J., and Crosby, W. I., (1997). The Yeast Two-Hybrid System.
Pp. 63-82. In: Hansen, E. and Harper, G., (ed), *Differentially Expressed Genes In
Plants*. Taylor and Francis Ltd, London.

Krisak, L., Strich, R., Winters, R. S., Hall, J. P., Mallory, M. J., Kreitzer, D., Tuan, R. S., et al. (1994). SMK1, a developmentally regulated MAP kinase, is required for spore wall assembly in *Saccharomyces cerevisiae*. *Genes & Development*, 8(18), 2151-2161.

Labaj, P. P., Lepar, G. G., Bardet, A. F., Kreil, G., & Kreil, D. P. (2010). Single amino acid repeats in signal peptides. *The FEBS Journal*, 277(15), 3147-57.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.

Liao, Y., Moir, R. D., & Willis, I. M. (2006). Interactions of Brf1 peptides with the tetratricopeptide repeat-containing subunit of TFIIC inhibit and promote preinitiation complex assembly. *Molecular and Cellular Biology*, 26(16), 5946-56.

Madden, K., Sheu, Y.-J., Baetz, K., Brenda Andrews, & Snyder, M. (1997). SBF Cell Cycle Regulator as a Target of the Yeast PKC-MAP Kinase Pathway. *Science*, 275(5307), 1781-1784.

Martin-Yken, H., Dagkessamanskaia, A., Basmaji, F., Lagorce, A., & Francois, J. (2003). The interaction of Slt2 MAP kinase with Knr4 is necessary for signalling through

- the cell wall integrity pathway in *Saccharomyces cerevisiae*. *Molecular Microbiology*, 49(1), 23-35.
- Mayer, B. J. (2001). SH3 domains: complexity in moderation. *Journal of Cell Science*, 114(Pt 7), 1253-63.
- Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, 38(Database issue), D204-10.
- Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots — a review of the protein–protein interface determinant amino-acid residues. *Proteins*, 68(4), 803-812.
- Neduva, V., & Russell, R. B. (2006). Peptides mediating interaction networks: new leads at last. *Current Opinion in Biotechnology*, 17(5), 465-471.
- Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein–protein interactions. *The EMBO Journal*, 22(14), 3486-3492.
- Nourry, C., Grant, S. G. N., & Borg, J.-P. (2003). PDZ domain proteins: plug and play! *Science's STKE : Signal Transduction Knowledge Environment*, 2003(179), RE7.

Ofran, Y., & Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS letters*, 544(1-3), 236-239.

Ofran, Y., & Rost, B. (2003). Analysing Six Types of Protein-Protein Interfaces. *Journal of Molecular Biology*, 325(2), 377-387.

Pawson, T., & Scott, J. D. (1997). Signaling Through Scaffold, Anchoring, and Adaptor Proteins. *Science*, 278(5346), 2075-2080.

Pawson, Tony, & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618), 445-452.

Pazos, F, Helmer-Citterich, M., Ausiello, G., & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4), 511-23.

Pazos, Florencio, & Valencia, A. (2002). In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs. *PROTEINS: Structure, Function, and Genetics*, 227(February 2001), 219-227.

Peña-Castillo, L., & Hughes, T. R. (2007). Why are there still over 1000 uncharacterized yeast genes? *Genetics*, 176(1), 7-14.

Petschnigg, J., Snider, J., & Stagljar, I. (2011). Interactive proteomics research technologies: recent applications and advances. *Current Opinion in Biotechnology*, 22(1), 50-8.

Petsko, G. A., and Ringe, D., Protein Structure and Function. Sunderland, MA: Sinauer Associates. (2004). 34.

Pitre, S, North, C., Alamgir, M., Jessulat, M., Chan, A., Luo, X., Green, J. R., et al. (2008). Global investigation of protein-protein interactions in yeast *Saccharomyces cerevisiae* using re-occurring short polypeptide sequences. *Nucleic Acids Research*, 36(13), 4286-94.

Pitre, Sylvain, Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., et al. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, 7, 365.

Plewczyński, D., & Ginalski, K. (2009). The interactome: predicting the protein-protein interactions in cells. *Cellular & Molecular Biology Letters*, 14(1), 1-22.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., et al. (2005). Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068), 679-84.

- Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., et al. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods (San Diego, Calif.)*, 24(3), 218-29.
- Robinson, L. C., Bradley, C., Bryan, J. D., Jerome, A., Kweon, Y., & Panek, H. R. (1999). The Yck2 yeast casein kinase 1 isoform shows cell cycle-specific localization to sites of polarized growth and is required for proper septin organization. *Molecular Biology of the Cell*, 10(4), 1077-92.
- Sambourg, L., & Thierry-Mieg, N. (2010). New insights into protein-protein interaction data lead to increased estimates of the *S. cerevisiae* interactome size. *BMC Bioinformatics*, 11(1), 605.
- Schaefer, M. H., Wanker, E. E., & Andrade-Navarro, M. A. (2012). Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Research*, 40(10), 4273-87.
- Schwer, B., & Gross, C. H. (1998). Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. *The EMBO Journal*, 17(7), 2086-94.
- Serebriiskii, I. G., & Golemis, E. A. (2000). Uses of lacZ to study gene function: evaluation of beta-galactosidase assays employed in the yeast two-hybrid system. *Analytical Biochemistry*, 285(1), 1-15.

- Sharifpoor, S., van Dyk, D., Costanzo, M., Baryshnikova, A., Friesen, H., Douglas, A. C., Youn, J.-Y., et al. (2012). Functional wiring of the yeast kinome revealed by global analysis of genetic network motifs. *Genome Research*, 22(4), 791-801.
- Skrabanek, L., Saini, H. K., Bader, G. D., & Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Molecular Biotechnology*, 38(1), 1-17.
- Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue), D535-9.
- Suter, B., Kittanakom, S., & Stagljar, I. (2008). Two-hybrid technologies in proteomics research. *Current Opinion in Biotechnology*, 19(4), 316-23.
- Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna Molina, M. M., Shames, I., Malitskaya, Y., et al. (2008). An in vivo map of the yeast protein interactome. *Science (New York, N.Y.)*, 320(5882), 1465-70.
- Torres, E. M., Dephoure, N., Panneerselvam, A., Tucker, C. M., Whittaker, C. A., Gygi, S.P., Dunham, M. J., et al. (2010). Identification of aneuploidy-tolerating mutations. *Cell*, 143(1), 71-83.

- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770), 623-7.
- Valencia, A., & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion In Structural Biology*, 12(3), 368-73.
- Walhout, A. J., & Vidal, M. (2001). High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods (San Diego, Calif.)*, 24(3), 297-306.
- Walther, T. C., Brickner, J. H., Aguilar, P. S., Bernales, S., Pantoja, C., & Walter, P. (2006). Eisosomes mark static sites of endocytosis. *Nature*, 439(7079), 998-1003.
- Wang, X., Hoekstra, M., DeMaggio, A., N Dhillon, Vancura, A., Kuret, J., Johnston, G. C., et al. (1996). Prenylated isoforms of yeast casein kinase I, including the novel Yck3p, suppress the *gcs1* blockage of cell proliferation from stationary phase. *Molecular and Cellular Biology*, 16(10), 5375-5385.
- Wojda, I., Bebelman, J.-P., Jakubowicz, T., & Siderius, M. (2007). Thermosensitivity of the *Saccharomyces cerevisiae* *gpp1 gpp2* double deletion strain can be reduced by overexpression of genes involved in cell wall maintenance. *Archives of Microbiology*, 188(2), 175-84.

- Wootton, J. C., & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17(2), 149-163.
- Xia, K., Fu, Z., Hou, L., & Han, J.-D. J. (2008). Impacts of protein-protein interaction domains on organism and network complexity. *Genome Research*, 18(9), 1500-8.
- Yan, C., & Dobbs, D. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20(Suppl. 1), 371-378.
- Yang, M., Wu, Z., & Fields, S. (1995). Protein-peptide interactions analyzed with the yeast two-hybrid system. *Nucleic Acids Research*, 23(7), 1152-1156.
- Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Science : A Publication of the Protein Society*, 3(5), 717-29.
- Zhou, H. X., & Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44(3), 336-43.