PROTEOME-SCALE PROTEIN-PROTEIN INTERACTION SITE PREDICTION AND NOVEL MOTIF DISCOVERY USING RE-OCCURRING POLYPEPTIDE SEQUENCES

by

Adam Amos-Binks

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

MASTER OF APPLIED SCIENCE

Biomedical Engineering

at

CARLETON UNIVERSITY

Ottawa, Ontario
January, 2010

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-91583-7

Our file Notre référence
ISBN: 978-0-494-91583-7

Table of Contents

List of Tables

List of Figures

Abstract

Protein-Protein interactions (PPIs) are an important area of research because most cellular processes are carried out by aggregates of interacting proteins. Identification of these aggregate complexes can provide insight into how the proteome is organized into functional units (A.-C. Gavin and et al., Nature, 415(6868):141-147, 2002) . Proteins can then be leveraged for such uses as therapeutic drug design where PPIs are often mediated or blocked by the drug.

This thesis is concerned with predicting binding sites that mediate a PPI by using re-occurring polypeptide sequences on a proteome scale, without the use of 3D structure data or gold-standard interaction site mediators. In this thesis, an algorithm has been developed to identify potential interaction sites, as well as a metric for evaluating the predictions. This research shows that interaction sites can be predicted from using re-occurring polypeptide sequences on a proteome scale for both human and yeast organisms.

Acknowledgements

A thank-you to all those who provided input, feedback, or support in the writing of this thesis.

Chapter 1

Introduction

1.1 Introduction

This chapter provides background information, outlines the problem, summarizes the work done to address it, and highlights the main contributions and results of this thesis.

1.2 Motivation

Proteins are organic compounds comprised of a sequence of amino acids, the definition of which is encoded in the genetic code of a gene, joined together by peptide bonds. Historically, proteins have been an important pillar of research within the biological and medical communities. The importance of understanding proteins and their characteristics is rooted in the complex, ubiquitous role proteins play in organisms and cellular function. Understanding disease development, metabolism, drug design, and DNA manipulation are some applications with which the knowledge about proteins can be applied.

In the past, research into proteins has relied almost exclusively on wet lab methods and techniques, where biological material is tested and analyzed using specialized equipment in a sterile environment. As classic wet lab techniques dig deeper and expand in scale, increasing amounts of knowledge are accumulated about proteins and the protein puzzle becomes more complex. While a great deal has been discovered about the role of specific proteins using wet lab techniques, the possibilities and field of application of proteins remains vast. Wet labs are extremely important for empirical, evidence based experimentation, but there are simply not enough wet lab resources to pursue all desired avenues of experimentation.

However, in recent years proteins have been garnering much interest in the engineering and computational sciences. Challenges with proteins and their properties, which have previously been constrained by the wet lab resources of biologists and medical researchers, offer interesting and challenging mathematical problems. These problems can often leverage state of the art engineering and computer science techniques and resources. Global or proteome (the entire set of known proteins) scale simulations can reveal re-occurring characteristics and intriguing results helping to prioritize and lead wet lab techniques in a more poignant direction.

Protein folding is a prime example of a wet lab problem leveraging engineering and computer science resources. Before a protein can perform any of its many possible functions, it must fold from a linear amino acid state to a three-dimensional native state. The folded three-dimensional structure can reveal much about the function of a protein, and it is thought that many diseases can be the result of mis-folded proteins [50]. Therefore, accurate solved protein folds can help researchers understand the development, spread, and implication of many diseases. Unfortunately, the exact fold of a protein can only be determined through expensive tests with specialized equipment such as x-ray crystallography and NMR (Nuclear Magnetic Resonance Spectroscopy) or through labor intensive processes such as energy landscape theory.

The FAH (Fold At Home) [32] project aims to increase the body of knowledge of known folded proteins without high costs in monetary or time resources by simulating protein folds using advanced computational methods. FAH uses a robust distributed computing algorithm to divide complex Molecular Dynamic (MD) simulations of protein folds among idle CPU cycles. In turn, the distributed nature of the FAH algorithm allows the FAH project to harness participants unused CPU cycles, from commodity x86 architectures found on desktops to high end GPUs (Graphical Processing Units) to modern gaming consoles, such as the Cell processor in the PS3. All that is required of FAH participants is an Internet connection and the installation of a small piece of software. Leveraging unused CPU cycles through the distributed algorithm allows the FAH project to obtain high computational throughput without thousands of hours from a dedicated, but costly, super-computing center. The FAH project has had many successes, and continues to focus efforts on problems related to

Alzheimer's Disease, Huntington's Disease, and Parkinson's Disease, among others.

While a renown method for the prediction of folded structures, the FAH project is not the only protein structure predicting method in use, but it may be the most popular. In fact, predicting a protein folding/structure is such an important and popular research activity that a bi-annual competition called CASP (Critical Assessment of Techniques for Protein Structure Prediction) is held to assess the state of the art in the field.

Protein-Protein interactions (PPIs) are also an important area of research because most protein cellular processes are carried out by aggregates of interacting proteins. Identification of these aggregate complexes can provide insight into how the proteome is organized into functional units [16] and can be leveraged for such uses as therapeutic drug design where PPIs are often mediated or blocked by the drug.

Y2H (Yeast Two Hybrid) is one of the first, and most popular, molecular biology methods used for direct recognition of PPIs. Y2H requires that two domains be present in order to perform an assay. The first is a DNA Binding Domain (DBD), and the second an Activation Domain (AD) for activating DNA transcription. To detect a PPI, the DBD is fused to one participant protein, while the AD to the other. If there is an interaction between the two hybrid proteins, a reporter gene is induced, and interaction detected. The Y2H technique is prone to missing true interactions (false negatives), as interacting proteins must be localized to the nucleus. Membrane proteins, for example, are unable to activate a reporter gene and will not register as an interaction, though they may have interacted. Y2H has been modernized and refined since its inception, and large scale Y2H (Yeast Two-Hybrid) [23, 53] assays have been carried out, and have led to understanding the underlying interactome, such as the identification of previously unclassified PPIs into a biological context [53].

TAP-tagging (Tandem Affinity Purification) is a generic procedure developed to purify and study proteins in their natural state in the cell [46]. In the tap-tagging process, fusion proteins are created by attaching the TAP-tag on the protein's chromosomal locus, then purified in a two-step process using Staphylococcus protein A and calmodulin beads separated by a tobacco etch virus (TEV) protease cleavage site [42]. After the two affinity purifications, the protein is examined for binding protein

partners. When compared with Y2H, TAP-tagging can identify a wider variety of complexes by using a more natural environment and physiological conditions (like pH requirements) [42]. Large-scale TAP-tagging experiments [30, 16] have yielded significant insights into cellular roles, assigning functional roles to proteins which previously had none, as well as assigning new roles to proteins to which roles had already been assigned.

However, while these traditional wet lab methods (Y2H and TAP-tagging) used for predicting PPIs are important, they are limited in scale (even the large scale experiments). Wet lab methods are also prone to false positives/negatives, contamination, and are expensive in terms of time, money, and lab resources. PPIs have many interesting characteristics and properties which translate into interesting mathematical problems, which has resulted in the development of many PPI prediction tools. Computational PPI methods have become an attractive alternative because of the response time, cost, and scale at which they operate. PPI prediction has become such a widespread research activity that a competition analogous to CASP is held for the assessment of prediction of interactions between proteins. CAPRI (Critical Assessment of Prediction of Interactions) [25] is an open competition taking place approximately every 6 months. Researchers are given the same proteins and attempt to dock the proteins together, a process simplified as two pieces of a puzzle fitting together, which is known as protein-protein docking.

The primary structure of a protein is at the core of predicting PPIs. A protein's primary structure is defined by the sequence of a gene (encoded in the genetic code), and is expressed as a unique sequence of the twenty amino acids which are joined together by peptide bonds. Any sequence of amino acids can also be referred to as a polypeptide sequence. When a protein's primary structure is discovered, it is said the protein has been "sequenced". The first protein to be sequenced was insulin in 1958. Co-ordinated efforts have yielded high quality repositories of protein sequences such as the NCBI (National Center For Biotechnology Information) RefSeq database contains over 9.3 million protein sequences.

Computational PPI prediction methods are able to produce large protein interaction networks, which map out interactions between protein pairs. Protein Interaction

Networks (PINs) can provide valuable insight into potential aggregates and complexes for use in drug design. However, even the most complete network is unable to identify the underlying mechanisms which mediate the interaction and binding of two proteins.

PPIs are mediated by the physical binding between small areas on the surface of a folded protein. Understanding the binding sites which facilitate PPIs can provide a great deal of insight into how a single protein fulfills its role in the proteome. When the amino acids in a binding site are altered by a mutation or some other manner (another partner), it can inhibit a protein's ability to participate in interactions. This can lead to disrupted signaling and lead to a change in phenotype for the organism, and it has been linked to several human diseases [28]. A comprehensive understanding of interaction sites can provide a great amount of insight for scientists and provide hypothesis-based starting points for drug design.

1.3 Statement of the Problem

The ability to predict interaction sites has lagged behind the ability to predict interactions themselves. Interaction site prediction methods based on protein structure require the solved three-dimensional structure and advanced docking techniques. The number of known three-dimensional structures is relatively small, and the running time required for most methods is substantial. These two shortcomings prevent proteome-scale prediction of interaction sites based on protein structural data.

A second class of methods based entirely on the sequence information available for proteins have shown it is possible to predict interaction sites from sequence alone. Sequence data for proteins is widely available, abundant, and methods using this information can be developed quickly. Predictions from sequence data are more suited for proteome-wide interaction site prediction.

Current sequence based methods require a list of motifs or domains known to mediate interactions. This will, at best, lead to the identification of known interacting subsequences. This is problematic as many interaction sites can simply be overlooked as they may not correspond to or be identified as an interacting motif or domain. Another limitation of using a list of gold standard motifs is the re-occurrence of

the same motif on a single sequence; current methods cannot identify which of the occurrences mediate the interaction when surrounding amino acids could prioritize them.

With the limited scale of current methods, the usefulness of a proteome-wide interaction site prediction is evident. A priori lists of mediating motifs are limited by their size and state of flux. Thus, the identification of new interaction mediating motifs over the entire proteome would be a contribution not found in any other method.

1.4 Contributions

The PIPE (Protein Interaction Prediction Engine) tool has shown to be adept at predicting novel PPIs with a 99.95% specificity on proteins from the H. sapien, and the yeast S. cerevisiae organisms [43]. The prediction method is based on reoccurring polypeptide sequences found in a database of known PPIs. Using the PIPE database of known PPIs and their co-occurring sequences, this thesis will demonstrate the ability to predict the binding site between two proteins.

The main contributions from this thesis include the algorithm used for identification of interaction sites from PIPE matrices, the distance metric to evaluate the accuracy of the predictions (which can be used by other future prediction methods to evaluate performance), and the identification of novel motifs in yeast and human proteins which may mediate interactions. The interaction site identification method presented in this thesis does not require protein structure information, or a list of known mediating motifs or domains, and the method can be applied at a scale not possible with other methods.

1.5 Organization of Thesis

Chapter 2 reviews literature on current methods to predict interaction sites. Chapter 3 provides an overview of PIPE, and the prediction method developed. Chapter 4 provides the details of the validation methodology, the distance measure (DM) developed, and evaluation of predicted sites. A discussion and interpretation of the

biological significance of the results from two datasets are discussed in Chapter 5. In Chapter 6, conclusions are made, contributions outlined, and direction for future work is identified.

Chapter 2

Previous work in Protein Interaction Site Prediction

2.1 Introduction

Amino acids are the building blocks of proteins and possess many properties which, when examined as subsequences of the primary structure, help determine expected behaviour of a protein. Signatures or longer common amino acid subsequences that have the ability to evolve, function, and exist independently of the protein are known as domains. Typically, short sequences of residues in a protein are the brokers of protein interactions. These short sequences form the contact interfaces between two interacting proteins and are referred to as the interaction site. Interaction sites can have any number of similar characteristics such as being geometrically complementary or contain motifs that are highly re-occurrent in other protein complexes.

Much work has been done in the field of interaction site prediction. There are a number of different methods which take into account multiple features known about interaction residues. A first successful automated attempt to predict interaction sites between proteins is the work by [21]. With the use of a Neural Network (see 2.2.2), trained on sequence profiles of neighbour residues and their solvent exposures as input, an accuracy of 70% (using exact interface matching as the metric) was achieved on a small dataset.

There are many emerging methods that make use of information for use in 3D structure analysis. A very popular method is 3D docking, where proteins are modeled in a three dimensional space. Computational geometry methods are then used to determine whether two or more proteins can fit together; if they can, an interaction is reported. However, the information required for 3D structural analysis is not widely available, as it requires x-ray crystallography and NMR to obtain the 3D information. Thus, it should be noted that more and more information is being added to repositories such as the Protein Data Bank (PDB). While 3D docking has

interesting problem sets, it is estimated that 40% of proteins cannot be modeled as legitimate structures [2], and further discussion of docking is beyond the scope of this paper.

With the scarcity of 3D protein data in mind, a second class of methods are emerging which make use of more widely available protein sequences and evolution data. The work of [39] and [45] has been unique in demonstrating the ability to predict PPI interaction sites solely on the basis of local sequence information. Building on this work, [54], [20] and [36] have developed methods which use advanced machine learning techniques with additional information (Section 2.4) to predict interaction sites.

2.2 Interaction Site Prediction From Primary Structure

The work of [57] classifies interaction site prediction methods into two classes: numerical value based and probabilistic. Each method relies on machine learning techniques, and only considers surface residues as interaction site candidates. This section discusses several techniques from the two groups. Neural networks (2.2.2) and Support Vector Machines (2.2.3) are discussed because of their importance in establishing the field, while Bayesian Networks (2.2.1) and Exact Binomial/Fisher's Test (2.2.5) are discussed as they are leading interaction site prediction methods most similarly associated with PIPE.

2.2.1 Probabilistic: Bayesian Networks

Bayesian networks are probabilistic models which detail the probabilistic dependence between a set of variables. They are often represented as directed acyclic graphs, where nodes are variables and edges are the directed dependencies between them. Bayesian networks are excellent tools for answering probabilistic queries about variables, as they are a complete model for variables and their relationships.

A motif refers to a defined sequence of elements within a protein's amino acid sequence. Motifs can be used to predict PPI and as a result obtain predicted interaction sites. InSite, a method developed by [54], is a significant contributor to interaction site prediction. It requires no knowledge of 3D protein structures, which many other

techniques need, and can make fine-grained predictions at the individual protein level using motif matching. There are several methods ([33], [19], [47]), which rely on the affinity between motifs for interaction site prediction and do not take into account other information for the participating proteins.

InSite takes input in the form of a dataset of prevalent motifs, a dataset of confirmed PPIs, and any other indirect evidence. The InSite authors list information on PPIs, motif-motif interactions, Gene Ontology (GO) annotations, and domain fusion as viable sources for indirect evidence. Using these inputs, InSite searches for high affinity motif pairs within the two proteins participating in the PPI to explain the interaction. A set of motif pairs, each with an estimated likelihood of producing the binding site is the output. This set of estimated likelihoods can then be combined with the indirect evidence to further refine whether certain motifs are legal binding sites. The selection of the highest probability, legally allowed motif is then used to form hypotheses of the form 'Motif M on Protein A bind to Protein B'.

The authors discuss the performance of InSite and it was shown to be significantly better at interaction site prediction [54] than other motif matching methods ([33], [19], [47]). The most significant result from the analysis was the error reduction of 43.7% and 19.8% for Pfam and Prosite datasets as well as 90% and 66% accuracy of the top 50 predictions. This indicates an ability to predict more accurate and a greater quantity of interaction sites than the other methods.

Of the methods reviewed, the InSite method shows the most promise in identifying interaction sites on a proteome-wide scale. The potential comes from integrating several simple but widely available data sources; however, there are two limiting factors. The first is that the training dataset requires the explicit identification of motifs and not just raw sequences. The consequence of this requirement is that no new legal motif will ever be discovered, since every interaction site identified can only ever be comprised of motifs, which have been provided in the training dataset. The second limitation is that all motifs are treated as potential binding sites, since many motifs are responsible only for protein characteristics such as structure. The replication of a single motif within a sequence is also problematic, as the InSite method is not able to determine which of the multiple instances of a motif would be the actual binding

site.

2.2.2 Numerical Value: Neural Networks

A Neural Network (NN) is a machine learning and data modeling tool that is often used to model complex relationships involving multiple inputs and multiple outputs. NNs are adaptive to their inputs and outputs, changing their connection weights based on the training data provided in the learning phase. The basis of a NN is from biological systems where a group of artificial neurons are interconnected to provide pathways for varying inputs which lead to differing outputs.

The work by [39] is an oft-cited article, since it is the first paper to use sequence information alone to predict interaction sites. The method uses a feed-forward NN with back propagation and momentum turn. The NN is then trained using windows of nine amino acids on consecutive residue sequences. [39] divided their dataset into thirds, two thirds of which was used as training data for the NN in order to predict the remaining third. The dataset was rotated through three times, so that each third would be used as the predicted portion.

Results from [39] include the observation that 98% of PPIs had, at minimum, one interacting residue consecutive in sequence. The authors identify a problem that is still relevant today: the false-positive rate is difficult to quantify. The main reason for this is the potential presence of other interaction sites not reported in the ground truth data. The authors adopt a very conservative true-positive measure by assuming that, for a given pair, there exists only the one site as identified in the ground truth dataset.

This landmark technique was crucial to proving interaction sites could be predicted using sequence information only. The downside of the technique is that the NN requires a training dataset, of which there are very few available, in comparison with what would be required for proteome wide experimentation.

2.2.3 Numerical Value: Support Vector Machines

Support Vector Machines (SVMs) is a supervised learning method used for classification of input into two classes. Data is viewed as a vector of p values in p-space

separated by a $p-1$ dimensional hyperplane. The goal is to maximize the distance between the closest points on either side of the hyperplane. SVMs are used by [45] and [29] to classify residues into interacting and non-interacting sites.

The work of [45] builds on the study of [39] by using SVMs (as opposed to NNs) with evolution information and protein residue conservation. Like [39], the goal is to identify interaction sites in PPIs without the use of structure information. The Real Value Evolution Trace (RVET), which uses phylogenetic trees to combine a grouping of related proteins and residue conservation, has been shown to increase the quality of predicted interaction sites [37].

SVMs were constructed using either residue composition, evolutionary information, or a combination in order to determine whether residues should be classified as interacting or non-interacting. The results obtained using the SVM classifier with only evolution information performed at approximately the same sensitivity, specificity, predicted positive value, accuracy, and correlation-coefficient as the SVM using composition (a reference implementation of the work by [39]), though it required 20 times less parameters. When combining the two SVM methods, better results were obtained, however, the authors expected the combined results to be higher. The authors theorize that due to the nature of protein 'hot spots', a small subset of the entire sequence that contributes the most energy to an interaction, results are skewed towards the hot spots ignoring the rest of the sequence. While the method was an improvement on the only existing work at the time, it still lags behind methods which use structure information for interaction site prediction.

2.2.4 Probabilistic: Conditional Random Fields

A Conditional Random Field (CRF) is a type of discriminative probabilistic model which is often used to process sequential data. A CRF is represented by an undirected graph where vertexes are random variables and edges are dependencies between them. The distribution of each random variable is to be conditioned on the sequential input. The work of [36] turns the prediction of protein interactions into a sequence labeling problem, and uses CRFs to process two protein sequences as inputs.

Protein structure data was extracted from the PDB, broken down into three features, and then fed into the CRF model. Spatially neighboring residue features and Accessible Surface Area (ASA) were used as the base feature set, and a Residue Conservation feature was used as an extended feature. Following the success of the base features in predicting interaction sites, the extended feature was added. The authors found that no increase in prediction accuracy was obtained by adding Residue Conservation.

The authors compared CRFs (using the base features) to other machine learning techniques; SVMs, NNs, and ME (Maximum Entropy) methods were used in their comparison. The authors found that the residue conservation feature did not contribute to the performance of the CRF and plan to investigate other features in hopes of determining the most effective ones. Other studies [39] used methods that found the residue conservation information to be useful. This may be the result of the chosen dataset, or, perhaps CRFs and sequence labeling are not techniques which make adequate use of all information.

2.2.5 Probabilistic: One-tailed Exact Binomial and Fishers Exact Tests

Using a binomial distribution model, the one-tailed Exact Binomial Test (EBT) compares the number of observed occurrences to the predicted number of occurrences. Fisher's Exact Test (FET) uses a geometric distribution model to detect rare observations in a population. These two statistical models can be used in bioinformatics to determine over-representation and under-representation, as well as rare observations of importance, in biological information.

Using the two statistical methods outlined above to determine the significance of motif observations, the work of [20] defines a set of Gold-Standard Positives (GSPs) and Gold Standard Negatives (GSNs) motif pairs. The assumption is that through rigorous statistical analysis, a set of high quality GSNs motif can be defined. With a set of high quality GSNs, the number of false-positive predictions can be reduced from the GSPs. Four datasets of GSNs were derived: one using a set of known non-interacting proteins; a second with proteins in differing sub-cellular compartments;

a third with differing sub cellular compartments as well as unrelated biological processes; and a final set with close sub cellular compartments but unrelated biological processes. The third dataset performed the best when evaluating the four datasets based on the motif pairs inferred by FET.

When compared with other methods, this method (termed Qvalue by the authors) was reported as more accurate when using the PPV (Predicted Positive Value) over datasets from DOMINO and IPfam. The authors attribute to both the rigorous statistical analysis performed and the small number of parameters that require tuning, making it less susceptible to noise from differing datasets, in comparison to the other methods.

This is the first use of a set of GSNs into a prediction method to reduce the false positive rate. Most prediction methods lack the ability to determine false positives. Thus, incorporating such a method would bring an increased level of confidence to interaction site prediction.

2.3 Interaction Site Prediction From Three Dimensional Docking

The success of such initiatives as the World Wide Protein Data Bank (wwPDB), which co-ordinates a massive solved 3D protein structure repository, have contributed a great deal to the accessibility to solved protein structures. The Fold-At-Home[32] has also contributed a great deal to knowledge of how proteins fold with over 70 publications since its inception in 2000. The increase in data and accessibility to it has led to protein structural information being used increasingly more in the prediction of physical binding sites. While it is usually necessary to have the structural information of a protein to predict its binding sites with other proteins, it has been shown that protein homologs physically bind in the same or similar ways [3].

Protein docking is the three-dimensional (3D) rendering process of a protein complex by attempting to fit together protein structures based on various physical, molecular and genomic properties. Typically each participant in the complex is rendered or folded individually, then surveyed for regions that match well (based on A Priori information about interaction sites) with other complex members. The first automated 3D docking algorithm was done by Wodak and Janin [26] in 1978, and there

are currently many ongoing research projects in this area.

The most recent docking and Molecular Dynamics (MD) simulation studies support protein complexes being formed in two stages [48]. The first stage is the initial collision encounter complex, identification of binding sites takes place through desolvation and burial of key hot spot anchors located in the center of the binding site. The stage is followed by a latching stage, where outlying binding residues may collapse by configuring their rotameric conformations into complementary arrangements [48].

There are two major challenges facing docking and MD simulations. The first is throughput. While the prices of commodity and high-performance computing equipment continues to fall, it can often still require upwards of 50 CPU days of compute time for a single experiment. Even when the running time is amortized over a cheap cluster, an experiment is often required to be run multiple times to identify multiple potential interaction sites. The ability to scale to proteome scale experimentation would take years in this case, during which time methods would have been refined and the results from the simulations may be irrelevant. The second major challenge is the number of false positives, which when combined with the number of hours required for a single experiment, negatively impacts the return on compute time used. Binding site prediction in general, whether by sequence or structure based, suffers a great deal from lack of validation data, as the cost of manual methods is so high.

Docking and MD simulations are starting to provide insights into how complexes are physically formed. In the future, docking and MD simulations are expected to become more tightly integrated with protein interaction site prediction software (such as the one developed in this thesis) in order to provide higher throughput and quality of predictions.

2.4 Interaction Site Features

Interaction site prediction methods rely on a set of features to deduce likely interaction site candidates. Zhou ([57]) organizes the most prevalent ones into the following five groups.

- *Sequence Conservation*: Interface residues (the components of an interaction site) occur in higher frequencies than non-interface residues, for either functional

or structural purposes.

- *Amino Acid Proportions*: Binding sites are often over represented by hydrophobic, aromatic and arginine residues, while underrepresented by others.
- *Secondary Structure*: Interaction sites favor β-strands while avoiding α-helices.
- *Solvent Accessibility*: Interface residues have shown to have higher solvent accessibility than their non-interface counterparts.
- *Side-Chain Conformation Entropy*: Interface residues have been shown to be less likely to sample alternative side-chain rotamers.

The authors of [57] suggest that of the features above, structurally related features are not susceptible to local conformational changes which often are associated with complex PPIs. In a study done by [15], the authors suggest that the identification of interacting residues is based on the hydrophobic moments, though a later study by [56] refutes that claim and reports a negative correlation between interaction sites and the hydrophobic moment.

2.4.1 Sequence Conservation

Sequence conservation, in the case of proteins, is the re-occurence of specific amino acid combinations across protein families. Sequence conservation has been used successfully in several studies in attempts to uncover underlying causes for protein interactions [45, 39, 21, 40].

A large number of domains, standalone functional units of a protein, have been found to be evolutionary conserved among proteins and are responsible for mediating interactions [24]. Domain-domain interactions have been studied using association rules and have been found to be responsible for a number of PPIs ([27], [49]). The use of association rules requires that the domains be conserved throughout a group of proteins, meaning there are a number domain-domain interactions identifiable through sequence conservation alone.

All studies reviewed focused on sequence conservation within specific protein families, a group of protein complexes, using homologs of proteins, or at a very small scale.

Sequence conservation holds great potential as input for proteome scale interaction site prediction because so many proteins have now been sequenced.

2.4.2 Hydropathy

It has long been thought that hydropathy is of major importance to interaction site prediction. Hydrophobic amino acids, such as isoleucine and valine, tend to be internal to a protein's three dimensional structure, whereas hydrophilic amino acids tend towards the surface. A study by [14] suggested the interaction site could be detected by the hydrophobic moment [11] only.

2.4.3 Charge

The charge of an amino acid, also known as the iso-electric point (IEP), is known to affect its participation in interaction sites [21, 8]. Both studies conclude that non-polar amino acids are more prevalent in interaction sites, conversely polar and charged amino acids are underrepresented.

2.5 Summary

While there are several methods which have shown the ability to produce high quality results, it is clear that there is no single method that will consistently predict the correct interaction site. The main reason is simply the lack of information. This is usually the case for structure-based and 3D docking methods. On a proteome-wide scale, this is also the case for sequence-based predictions, as the need for a substantial size training dataset of gold standard positives becomes more important. Methods which aggregate sequence, structure, and other features (see Section 2.4), such as the study by [54], show the greatest potential for large-scale prediction.

Chapter 3

Interaction Site Prediction from Re-occurring Polypeptide Sequences

3.1 Introduction

This chapter will provide on overview of elements and terminology essential to protein interactions. This section introduces the limitations with current interaction site prediction methods, describes an overview of the PIPE method, outlines a proposed solution, and identifies the significance and originality of the solution.

3.2 Predicting Interaction Sites From Sequences

Interaction site identification is an area where few proteome scale experiments, either predicted or wet lab, have been done. Predicting interaction sites from sequence has shown to be possible by several studies [39, 47, 21]. The underlying principles of sequence based predictors is the conservation of sub-sequences (often referred to as domains and motifs) from previously identified interaction sites. Yet, sequence based predictors may not be able to capture all of the subtleties between two proteins. For example, some known binding sites are non-contiguous in a protein's sequence. The goal of sequence-based predictors is to identify all the interaction sites that can possibly be identified from sequence alone.

3.3 Limitations of Current Methods

While the significance of the results from the methods surveyed in Section 2 is not to be ignored, there are limitations inherent in all the methods.

Structure-based predictions suffer from lack of input data, namely an abundance of three-dimensional solved protein structures. This lack of input data limits structure-based prediction in terms of proteome scale prediction. Structure-based predictions

are often time-consuming and require complex algorithms in order to reach a solution.

Sequence-based predictions use widely available amino acid sequences of proteins. However, most methods also require a list of gold-standard mediating motifs and/or domains in order to identify potential interaction sites. Different statistical, graph-theoretic and machine learning techniques are then applied to sequences and a predicted site is determined. The major limitations of these approaches, is that each method is constrained to predicting a gold standard motif or domain. There exists no possibility to predict interactions which are mediated by novel motifs/domains. Additionally, many motifs and domains can re-occur multiple time in a sequence. This can lead to ambiguity when predicting motifs, as it is not clear which of the multiple motifs is actually participating in the binding between two proteins.

InSite is currently the leader in predicting protein interaction binding sites on a proteome wide scale. However, in can only identify a motif on one side of the interaction.

The validation of predicted sites is also an area that can be improved. The quantity and quality of lab-confirmed interaction sites needs to be improved in order to provide more accurate evaluations of tools and to facilitate comparisons between them. Increasing the quantity of the results available will add statistical significance to predicted results. Precision improvements with respect to the location of interaction sites, as well as the identification of multiple sites will add more confidence to the predictions.

3.4 Protein Interaction Prediction Engine: PIPE

Using a dataset of known, confirmed PPIs and their amino acid sequences (henceforth referred to as the PIPE DB), the PIPE method determines the probability that a given pair, Protein A and Protein B, will interact. The calculations are based on a sliding window over a target protein's sequence and the re-occurrence of the window contents in the PIPE DB.

The PIPE method has four major steps which are outlined in Figure 3.1 (reproduced with permission from [43]). The first of which is to transform all confirmed

PPIs into a graph representation. The source of the PPIs can be from Tandem Affinity Purification (TAP) Tagging, Yeast Two Hybrid (Y2H), or any other experimental technique. The second step is to use a sliding window of twenty amino acids over one of the input sequences (sequence A) and to match the window to a sequence in the graph of confirmed interacting proteins. Once a match is discovered in the graph for sequence A, step 3 begins. Step 3 compares the adjacent neighbours, determined from the graph, of the match (to sequence A) to the second input sequence (sequence B), using a sliding window. If there is match, the result matrix is updated with the number of matching proteins. The above steps are repeated for the entire sequence of A, and once completed, the final step of visualizing the matrix is performed.

The PIPE method identifies frequently occurring known interacting subsequences between a candidate pair of proteins. The output matrix (henceforth referred to as the PIPE matrix) columns represent the sliding window of amino acids over sequence A, and the rows represent the windows for sequence B. The values in the matrix cells represent of the frequency at which the respective polypeptide sequences from protein A and B co-occur. The highest values in the matrix are indicative of the most frequently occurring polypeptide sequences (henceforth referred to as peaks), and offer the most likely candidates for interaction sites. This thesis investigates whether or not the peaks from the PIPE matrix are indicative of the actual binding locations for given PPIs. The PIPE method has shown to be adept at predicting yeast PPIs on a proteome scale [44], as well as human PPIs (publication forthcoming).

The terms 'binding location' and 'interaction site' are used synonymously. An interaction site is defined as a range, comprised of a start and end point, in the amino acid sequence for both proteins involved in a PPI. Three peaks from the PIPE matrix are identified and considered for analysis. The basis for this is that to identify more than three visually would result in a very ambiguous or low confidence interaction site.

3.5 Using PIPE To Predict Interaction Sites

As discussed in 3.4, the PIPE method identifies the co-occurrence of amino acid subsequences of length 20. For any 2 protein sequences, all possible combinations of

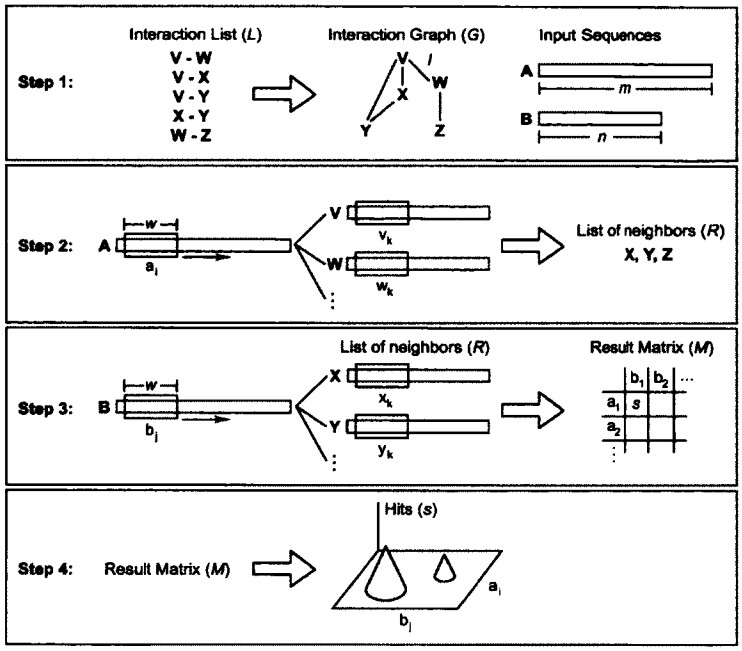

Figure 3.1: PIPE method for predicting PPIs

subsequences are found, and are represented by a frequency matrix. The rows and columns represent a subsequence of 20 amino acids from each of the proteins in question, the values in each position of the matrix represent the frequency at which the two subsequences are observed in a database of known interacting proteins. Based on the assumption that interaction sites are highly conserved [18], the highest values in the matrix (indicating the highest co-occurring subsequences) would represent potential interaction sites between the two proteins in question.

3.5.1 Advantages of PIPE method

Using the PIPE method to identify areas with highly conserved sequences allows the selection of local maxima as the interaction site. No training data beyond a database of known interacting proteins and their sequences is required, since both are abundantly available. Using large amounts of widely available data eliminates the dependence on gold standard mediating motifs and domains all other prediction methods use. The removal of this dependence permits interaction site prediction at

a proteome-scale.

3.5.2 Basic Peak Selection Algorithm

When the scoring matrix from the PIPE method is visualized in three-dimensional space, the most conserved subsequences are characterized by smooth 'peaks'. Henceforth, the term 'peak' will be used interchangeably with interaction site and binding site. This chapter outlines the basic peak detection algorithm, identifies limitations, and then details the steps taken to address them.

While it is simple to visually identify the peaks from the PIPE matrix, it is unfeasible to do this for large scale experimentation. The level of difficulty in determining the peaks from the PIPE matrix can range from trivial (Figure 3.2), to ambiguous (Figure 3.3), to undefined (Figure 3.4).

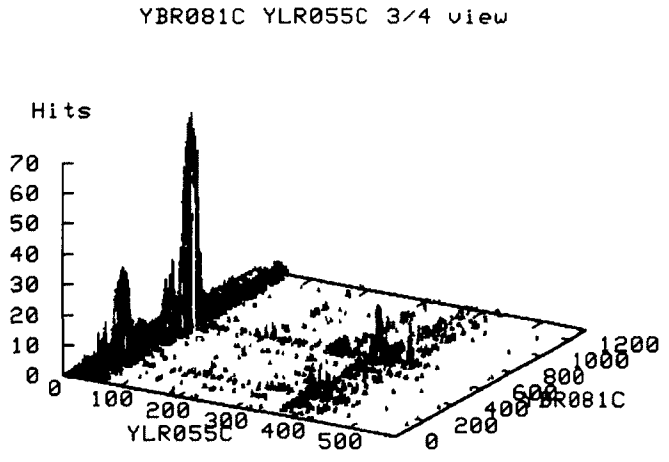

Figure 3.2: Trivial Peak in PIPE Matrix

The goal of the algorithm is to predict, from the PIPE matrix, a subsequence in both proteins which participate in the interaction. The first step to identify a peak in the matrix is to select the highest point in the matrix, N_{ij}. From the highest point N_{ij}, the algorithm will recursively step to the adjacent cells N_{i-1j}, N_{i+1j}, N_{ij+1}, N_{ij-1}, until the values in the cells are 0. When the algorithm completes its recursive

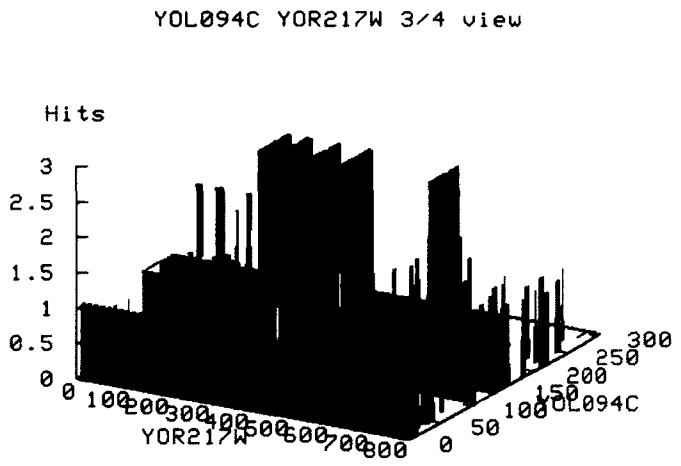

Figure 3.3: Ambiguous Peak in PIPE Matrix

stepping, the window size is added on to the range, and then each protein subsequence is reported.

A simplified pseudo-code of the basic peak selection algorithm:

```
while !done
{
        startPoint = getHighestPoint( pipeMatrix );
        while (walkLeft or walkRight or walkAbove or walkBelow)
        {
                area = area + numWalkableDirections;
        }
        if (area > minArea)
                done = true;
}
```

3.5.3 Basic Peak Selection Limitations

Simply identifying the highest point in the PIPE matrix followed by recursive stepping did not yield the desired peaks. To account for the variability in the PIPE

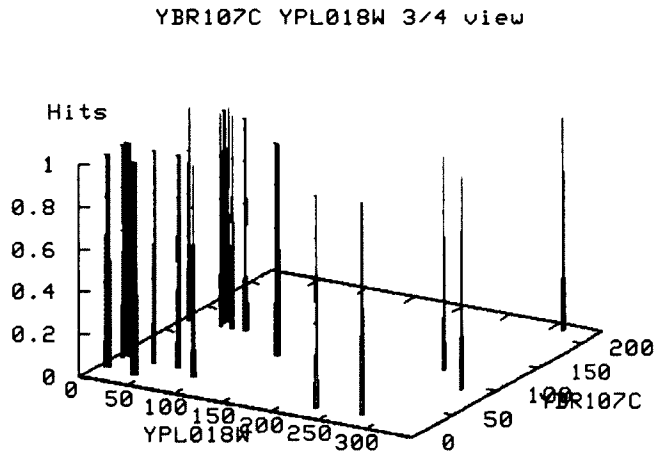

Figure 3.4: Undefined Peak Observation

matrix landscapes, two key issues needed to be addressed to ensure correct peak identification.

The first issue is to address spurious 'spikes', single high-value points in a matrix, that need to be screened out. Spikes occur from an amino acid sequence that has, by a random chance, matched a frequently-occurring sequence. The PIPE method uses a window of twenty amino acids to compute the frequency matrix. Frequent, re-occurring amino acid sequences will appear as a smooth curves using this method, as a re-occurring subsequence will gradually be shifted out one amino acid at a time, as the window moves forward on the sequence. Conversely, spikes have matched a re-occurring sequence of twenty amino acids, but as soon as the window shifts by a single amino acid, the subsequence does not match any frequent observations in the database. This is observed as a steep drop off when the frequency matrix is rendered in three dimensions. Spikes are not indicative of conserved subsequence and do not have the characteristic smooth curve or gentle slope observed in higher confidence peaks. Identifying spikes involves comparing the difference in height of a spike's adjacent cells in the matrix to the spike itself; the larger the difference, the higher the likelihood it is a spike.

Narrow, medium-height frequencies spanning the entire length of a protein (henceforth referred to as walls) are often observed where there is a frequently-observed sequence in one of the proteins in the PPI, but none in the other side. The effect of including a wall in an interaction site prediction is a reduction in accuracy and precision of the prediction, due to the predicted site spanning the entire length of the protein. Observing a wall in the PIPE matrix is not indicative of an interaction site, and only reduces the accuracy of PIPE's prediction. Walls can be excluded from interaction sites by requiring a minimum height, relative to a peak's height, in order to be included in the predicted interaction site.

3.5.4 Web Portal

The peak selection algorithm, interaction site features, and domain identification have been integrated into the existing PIPE portal and are available for use in the PIPE3b Web Portal, http://cgmlab.carleton.ca/PIPE3b/.

3.6 Summary

There exists a significant amount of widely available data for proteins, from both sequence data databases such as SGD (Saccharomyces Genome Database) and Entrez (a NCBI database), as well as PPI information from databases such as DIP (Database of Interacting Proteins), MIPS (Mammalian Protein-Protein Interaction Database), and BIND (Biomolecular INteraction Network Database). Improving the PIPE method to accurately predict and scale interaction site prediction, by leveraging the available data, is of great use as an alternative or precursor to wet lab experiments.

Chapter 4

Experimental Validation

4.1 Introduction

This chapter provides details of the validation datasets, and the metrics used to configure the parameters in the interaction site prediction. A discussion of the identification of novel motifs within predicted interaction sites using the Prosite database follows.

4.2 Interaction Site Validation Datasets

In order to evaluate PIPE's accuracy for predicting interaction sites, ground truth datasets were required. Using the DOMINO (Domain Peptide Interaction Database, http://mint.bio.uniroma2.it/domino/) open-access protein interaction repository, a dataset of 464 two-sided lab-confirmed interaction sites of PPIs from yeast (*S. cerevisiae*) , and a dataset of 506 two-sided lab-confirmed interaction sites for PPIs from human (*H. sapien*) were obtained. The interactions in DOMINO are identified with gene names, and need to be translated into SwissProt names, which is what the PIPE database uses. Multiple gene names and IDs can map to a single Swissprot ID, which is why there is a reduction in the DOMINO interactions and the number of interactions in the PIPE database.

Both the human and yeast datasets are complete with start and end sequence positions for each participant in the PPI which have been derived from lab-deletion analysis. Lab deletion analysis is a wet lab technique which confirms an interaction between two proteins, removes half of a protein and attempts to confirm the interaction once again. This process is repeated until the two proteins no longer interact, at which time the last sequences removed are assumed to have been responsible for

Protein A	Protein B	1-Sided	2-Sided	2-Sided in PIPE	Peak GT 10
H. sapien	H. sapien	785	506	423	363
H. sapien	Other	225	147	N/A	N/A
S. cerevisiae	S. cerevisiae	850	464	265	176
S. cerevisiae	Other	0	0	N/A	N/A

Table 4.1: Protein-Protein interactions with site locations extracted from DOMINO

the interaction. This process is time consuming, expensive and prone to contamination. The lack of lab-confirmed validation data only underscores the need for more reliable predictions, as wet lab techniques have not been popular areas of research. It is known that PPIs often involve multiple sites in a single interaction. A protein's interaction sites vary in different PPIs depending on its partner protein.Though a set of uniquely confirmed interaction sites for each PPI is desirable, other search attempts were unsuccessful, and the dataset from DOMINO was used.

The validation data used throughout the experiments was based on two-sided confirmed sites (both proteins had start and end ranges). One sided data (only one of the protein's start and and ranges were reported) for PPIs is available from DOMINO, but not useful, as this thesis is concerned with co-occurrence and identifying interaction sites on both proteins. Table 4.1 displays the number of interactions for both H. sapien and S. cerevisiae obtained from DOMINO.

4.3 Interaction Site Validation Methodology

4.3.1 Distance Measure Calculation

As discussed above, the DOMINO dataset does not contain the granularity required to identify multiple interaction sites in a single PPI, or differing participant sites of a protein present in multiple interactions. Any binary (Yes/No) evaluation of a predicted interaction site would be assuming that the PPI had only one site, and that it was used in all other PPIs for that protein. This is simply not the case, and so a more accurate way to measure, given by a continuous scoring metric, was required.

In the literature review, no existing continuous evaluation or comparative metric was discovered. Due to the nature of motif matching and docking problem spaces, current prediction techniques are only evaluated using binary comparisons and exact

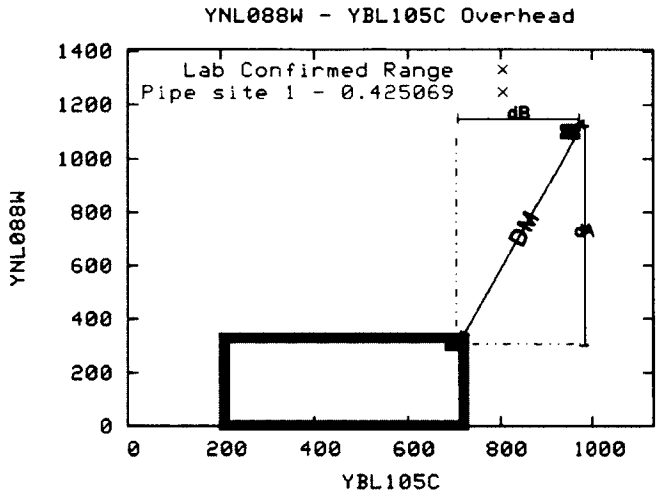

Figure 4.1: Distance Measure (DM)

matching. Using non-exact matching, as in PIPE, requires the development of a method to measure the relative distance and accuracy of a predicted site to a lab-confirmed site.

The goal of the distance measure (DM) developed is to evaluate the relative distance from PIPE's predicted site (or any other site) to a lab-confirmed site. The dataset given by DOMINO has only 2 dimensions; therefore, the PIPE DM will only consider the difference in area between DOMINO and PIPE interaction sites, and not the difference in volume (this would require 3 dimensions from DOMINO). When viewing the confirmed and predicted sites on a Cartesian plane, Figure 4.1, with the X,Y axis corresponding to the start and end ranges of protein A and protein B respectively. Because a range is being considered (effectively 4 points) and not a single point, an additional criteria is enforced; the furthest point in the predicted site is required to be inside the confirmed site. In Figure 4.1, this requires measuring the distance from the furthest point (X,Y) away in the predicted site from the closest point (X,Y) in the confirmed site. This is represented by the equation in 4.1:

$$dA = max[(startA_{labSite} - startA_{pipeSite}), (endA_{pipeSite} - endA_{labSite})] \qquad (4.1)$$

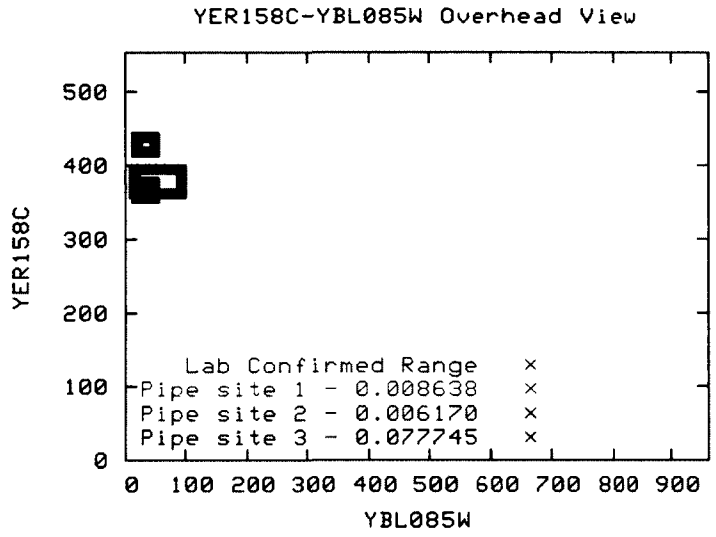

Figure 4.2: Overhead view of PIPE predicted interaction sites with DM

The distance needs to be calculated for each protein in the PPI, dA and dB (dA is only shown for simplicity). Some protein sequences are significantly larger than others; to account for this dA and dB are normalized by each proteins sequence length, and the equation becomes 4.2:

$$dA = \frac{max[(startA_{labSite} - startA_{pipeSite}), (endA_{pipeSite} - endA_{labSite})]}{proteinA_{length}} \tag{4.2}$$

Before combining the contributions from both Protein A and Protein B of the PPI, the vectors are squared, allowing the dominant vector to influence the score more significantly. The distance measure equation is defined in 4.3.

$$DM = \frac{\sqrt{dA^2 + dB^2}}{\sqrt{2}} \tag{4.3}$$

The division by $\sqrt{2}$ is to ensure a score lies somewhere between 0.0 and 1.0, where 0.0 is a perfect prediction (prediction site is within the confirmed site), and 1.0 is the furthest away possible. The goal behind the DM is to favor smaller more accurate prediction sites and penalize longer sites.

4.3.2 Addressing Limitations of Basic Peak Selection

In order to address the limitations described in Section 3.5.3, two parameters are introduced in the algorithm to refine peak selection. The first parameter is a measure and requirement of the quality of a peak's members. By introducing a minimum height restriction for each point from the matrix in the interaction site, a higher precision was achieved in the predicted range. Problem areas, such as walls, which would significantly alter PIPE's otherwise precise predicted interaction site, are ignored. The introduction of a second parameter is a measure and requirement of quantity. Requiring that a peak has a minimum number of adjacent neighbours of a specified height excludes random spikes from consideration as potential interaction sites.

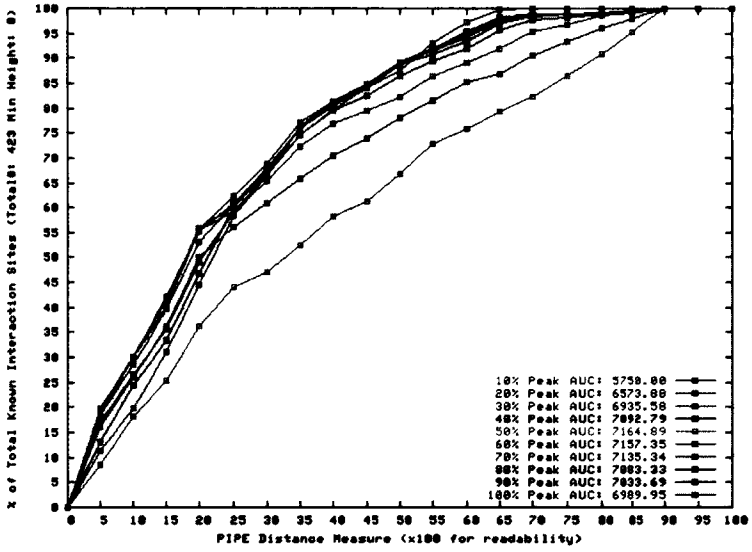

Figure 4.3: H. Sapien DM using variable Peak height %

Previous work in [43], using a median filter to identify spikes, found that it was only necessary to verify the neighbours immediately adjacent to a point in question. Building on this work, the minimum number of adjacent points required for verifying the legitimacy of a peak was set to five (the point itself, and the points directly to the left, right, above, and below). While this was merely a starting requirement, peaks were correctly selected in all interactions predicted by PIPE for both human and yeast datasets, with no random spikes ever being selected as interaction sites.

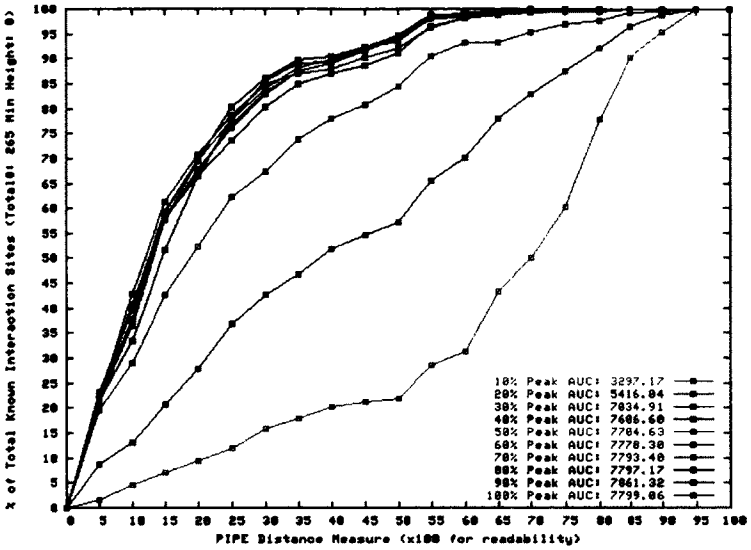

Figure 4.4: S. cerevisiae DM using variable Peak height %

With these successful results, no further tuning was done, and the minimum number of neighbours was set at four.

The second parameter, a neighbour quantity requirement, is tightly coupled to the first parameter and plays a major role in the accuracy of the predicted sites of PIPE. Figures 4.3 and 4.4 show the increase in accuracy of the DM scores for PIPE predicted sites as the minimum neighbour height quantity is increased. The height was increased by an interval of 10 from 10% of the peak height to 100%. It can be seen that as the peak height approaches 100%, more lab-confirmed results are excluded because of the constraints, however some DM scores become marginally more accurate. When using a 50% peak height in human data, the AUC (Area Under Curve) was the greatest, while in yeast 70% was the highest. In this case, the AUC can be somewhat misleading, as the desired outcome is that of having a very high steep slope at the beginining of the chart. A steep slope indicates that scores are skewed with favorable DM scores and more accurate predictions obtained. Considering that in both yeast and human using 50% peak provides both a leading AUC score and a steep slope, further experimentation uses a 50% peak height as its default.

With the above parameters configured for the selection of the highest peaks, and

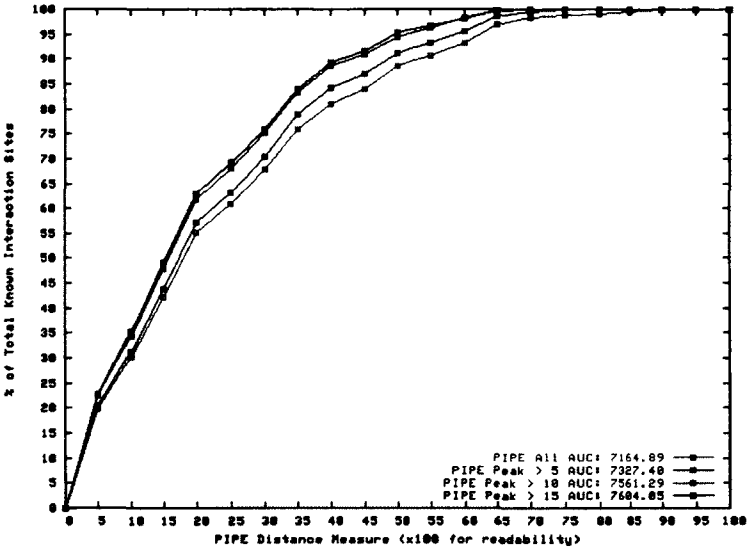

Figure 4.5: H. Sapien DM using overall min. peak height

avoidance of walls and spikes, the best interaction site candidates are chosen. However, some of the predicted interactions have very few re-occurring sequences. This can be due to lack of similar protein structures in the PIPE database or spurious lab results (perhaps an experiment was contaminated). Irrespective of the reason, the resulting matrix of these interactions has a very flat landscape and no obvious area of interaction to select. Observing 4.5, and 4.6, it can be seen by the AUC increase, and the greater height at the beginning of the curve, that the accuracy of the peak-finding algorithm is improved significantly if interacting pairs without a matrix point of 10 or higher are removed from the analysis. This observation is also consistent with previous research indicating that the probability of two random polypeptide sequences to re-occur more than 10 times is 10^{-6} [43]. Avoiding the prediction of an interaction site on a PPI with a peak height height less than 10 results in the loss of 14% and 34% in human and yeast validation data respecitvely (see Table 4.1).

4.3.3 Size Measure Calculation

The size measure was introduced in order to identify the percentage of a peak that should be taken into account when identifying the interaction range. By taking into

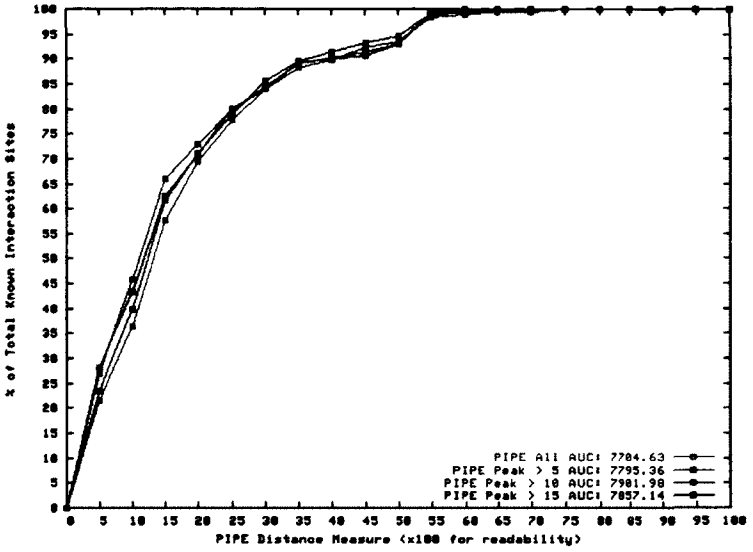

Figure 4.6: S. cerevisiae DM using overall min. peak height

account the size of the sites in our validation dataset, we can determine what an acceptable peak size estimate would be on a large scale.

The size measure for an interaction between two proteins predicted by PIPE is defined in Equation 4.4

$$SM = \frac{|LabLength_A - PIPELength_A| - |LabLength_B - PIPELength_B|}{LabLength_A + LabLength_B} \quad (4.4)$$

In both 4.7 and 4.8, the CDF (Cumulative Distribution Function) curve is steepest (meaning a higher density of better scoring sites) and the AUC total is highest, when using 50% of the peak height. This indicates that using 50% of a peak's height for the predicted interaction site will yield a size that will most closely match observed interaction site sizes.

4.3.4 Random Site Comparison

The literature revealed that existing techniques either do not provide the sequence ranges of their predicted interactions or lack the required information (3D structures) to predict desired interaction sites. Conversions of predicted interaction sites would also not be suffice as predictions on individual proteins are only provided on one of

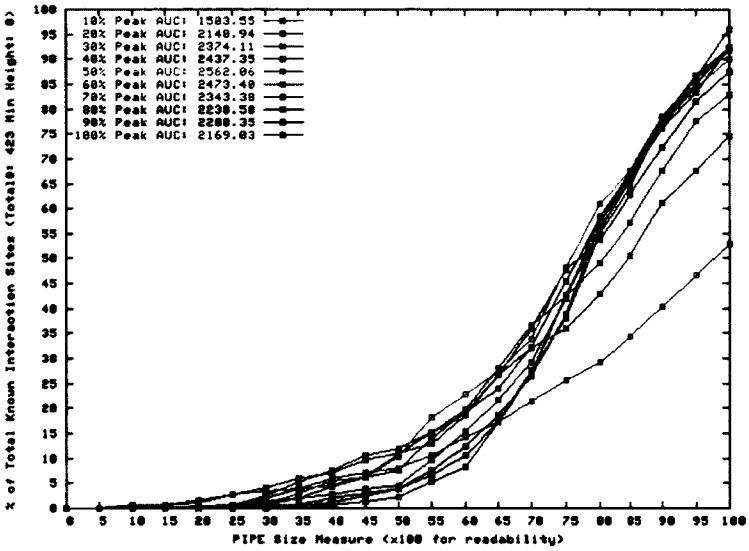

Figure 4.7: PIPE Site SM (H. Sapien)

the proteins in the PPI, or the predictions are coarse grained and at the protein family level, not the individual protein level. The lack of available tools to compare against PIPE required the establishment of a baseline performance measure. Using a random number generator, random interaction sites were chosen by selecting a random start point within a protein's sequence, and then selecting a point greater than the start point as the end of an interaction range. This process was performed for both proteins in the PPI, and then the equation 4.3.1 was calculated to obtain the distance measure (DM). Figures 4.9 and 4.10 show the cumulative distribution using DM scores for PIPE against the random site selection described above. While a random distribution model is used, it would be possible to use a more sophisticated random model using the interaction site size, and sequence locations from the validation data.

With no results from other PPI site prediction tools readily available, DM scores are baselined against random site selections. Figures 4.9 and 4.10 show the best and worst case sites for each PPI. Observe that the AUC worst case of PIPE is still significantly better than the AUC of the best case of random site selection in both datasets. While it does not provide a concrete answer about the accuracy of using the PIPE matrix for interaction site detection, it does provide reason to conduct further

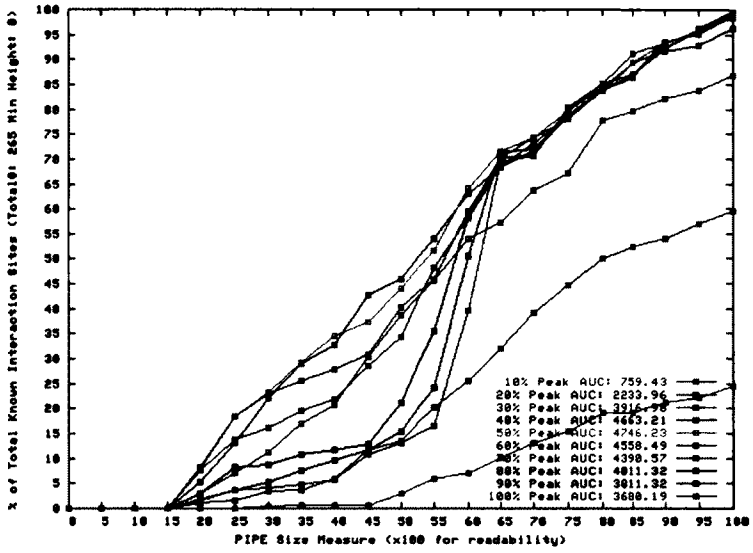

Figure 4.8: PIPE Site SM (S. cerevisiae)

research and determine which interactions are best predicted by this approach.

Figure 4.11 demonstrates the difference in the quality of interaction site predictions between the two species, and the random predictions. Most importantly is that both human and yeast are significantly better than choosing random sites. The overall AUC for the yeast dataset is greater, though only by 4.4%, which can be attributed to the size of the interaction graphs, from which the re-occurent polypeptide sequences are derived, and, in turn, construct the PIPE matrices. The underlying PIPE DB which is the source of the predictions made for each species comprises of approximately 36,000 PPIs (edges) for the yeast PIN (Protein Interaction Network), while the human PIN has approximately 40,000 PPIs. Both the yeast and human PINs are using the latest version of the PIPE DB which is, at the time of the writing of this thesis, not yet published. It is worthwhile observing that while the AUC is greater for yeast than human, the best predictions for both species, those with a DM score of less than 20, are very consistent with one another. Of the human PIPE predicted interaction sites, 63% had a DM score of less than 20, while 73% of PIPE predicted yeast interaction sites.

The parameters are tuned to ensure the peaks selected from the PIPE matrix are

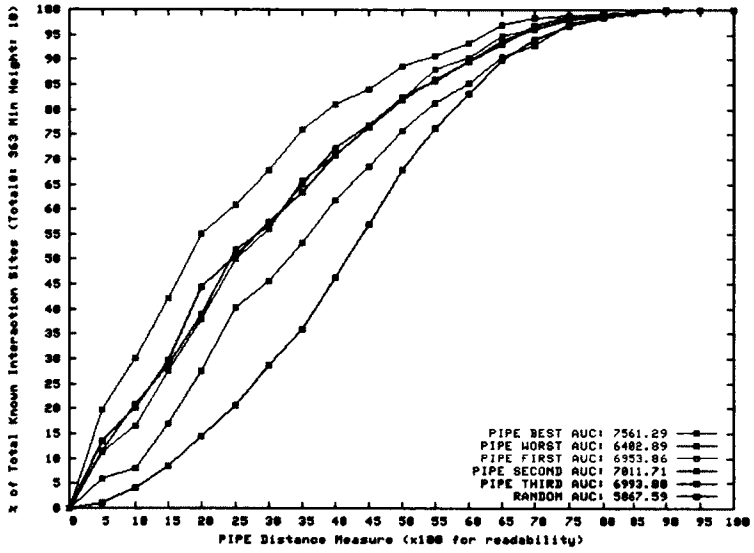

Figure 4.9: PIPE Site DM (H. Sapien) vs Random Site DM

the peaks one would identify visually. The three highest peaks are chosen by the method described in 3.5.2, more than three candidate peaks would likely indicate a very noisy matrix and ambiguity as to which peak would be the best candidate. Figures 4.9 and 4.9, demonstrates there is not a significant difference in total AUC or slope between the first, second, or third peak selected.

4.3.5 Balance In Validation Data

There are some duplicate participants in the dataset of lab-confirmed interaction sites. Of the 265 confirmed interaction sites in the yeast dataset, there are 140 unique participants, with one protein involved in 72 PPIs. The human dataset contains 423 PPIs with 364 unique protein participants, and the single highest protein is involved in 58 PPIs. Duplication of participants in PPIs could potentially skew the DM scores unfavorably in the case of a consistently poorly predicted protein or favorably if it is a consistently 'easy' prediction. In this case, because of the relatively good predicted DM scores for most proteins, consistently poor DM scores could be indicative of either a mistake in the lab, the reuse of a single interaction site where multiple in fact exist, or an interaction site that changes in the different PPIs a protein is involved

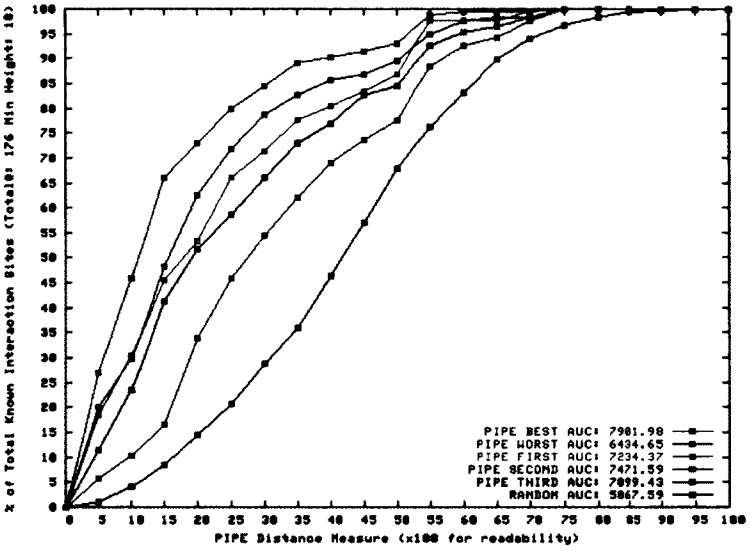

Figure 4.10: PIPE Site DM (S. cerevisiae) vs Random Site DM

Protein	Freq.	Avg. DM contribution
P04626	58	0.233
P00533	48	0.097
P21860	35	0.240
Q13444	21	0.012
P62993	13	0.022
Q15303	11	0.210
O95400	10	0.447

Table 4.2: Most Frequent Proteins in Lab Confirmed Interaction Sites (Human).

in. In order to determine if there was a group of proteins skewing the DM scores, the individual contributions, dA and dB were measured for the first peak found in the PPI.

Tables 4.2 and 4.3 show the most frequently occurring proteins in the lab-confirmed dataset: the protein name in column one; the number of occurrences in column two; and the average DM score in column three. In Table 4.3 it can be seen that for the protein YHR016C 4.4.4, the average DM contribution is an impressive 0.069. Section 4.3 discusses the domain associated with the prediction and lab-confirmed interaction site.

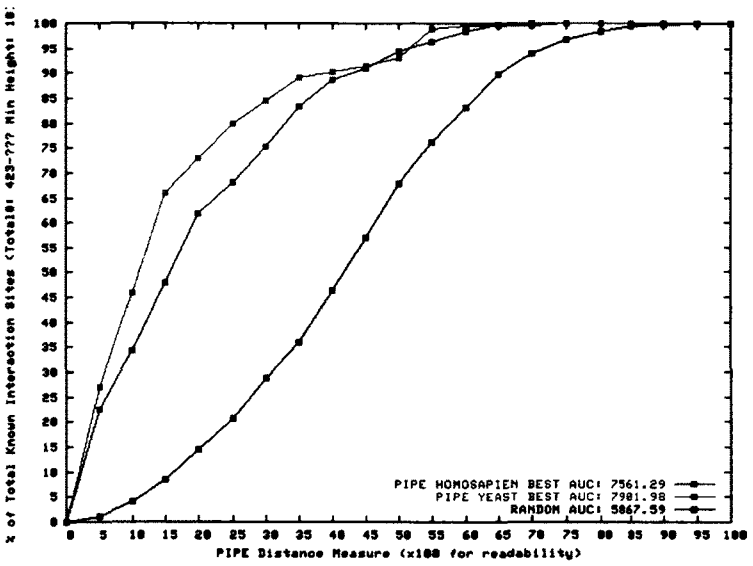

Figure 4.11: PIPE DM (S. cerevisiae) , PIPE DM (H. Sapien) vs Random DM

Protein	Freq.	Avg. DM contribution
YFR024C-A	72	0.149
YHR016C	48	0.069
YDR388W	33	0.119
YCR088W	25	0.128
YMR109W	24	0.040
YER118C	17	0.150
YBL007C	13	0.198
YBL085W	13	0.138

Table 4.3: Most Frequent Proteins in Lab Confirmed Interaction Sites (Yeast).

4.4 Interaction Site Domains

While predicting the interaction site subsequence is itself a challenging problem, mapping the subsequence to existing protein subsequence patterns is of great interest to the biology community. Domains provide context for the potential functional uses of a peptide sequence. Identifying domain pairs which were previously not known to interact, or ones that are known to mediate interaction, helps determine the confidence of a prediction. The co-occurrence of certain domains can explain why two proteins are known to interact. Conversely, predicted sites which are not mapped to

existing domains are also of great interest, as these subsequences hold potential as novel, undiscovered domains or motifs.

4.4.1 Prosite Membership

The Prosite open access database is a repository of protein descriptions, families and functional sites along with patterns and profiles used to identify them [22]. The ScanProsite tool [9] is used to identify domains and motifs in an amino acid sequence without A Priori knowledge of the protein the sequence was from. ScanProsite is available for download for use on a large scale. The use of ScanProsite provides biological significance and context to the predictions made in this thesis. Using the ScanProsite tool, the predicted interaction sites sequences were provided as input, and domains contained within the sequence were the output. ScanProsite contains options to perform partial matching on domains. However, a conservative approach was adopted and used throughout the experiments which requires a domain to be completely inside a predicted interaction site.

Domain identification within interaction is significant for two reasons. Identifying Prosite domains confirms that the peak selection algorithm and PIPE method is, in fact, identifying potential interaction sites, and not just low-complexity regions within a protein's primary structure. Secondly, the re-occurrence of domains from both participants in the PPI can characterize what types of interactions the method can accurately predict.

4.4.2 Absence Of Domains Within Predicted Sites

Interestingly in Table 4.2, the protein Q13444 has an extremely accurate DM contribution of 0.012. However, no Prosite domain is associated with the confirmed interaction range or the predicted range. This makes this amino acid subsequence of interest as a new domain or motif, as it is conserved on a proteome scale, and an excellent candidate for a wet lab experiment confirming its functional role.

4.4.3 Co-occurring Domains Within Predicted Sites

Domain information from predicted interaction sites can be used to provide insight into which interaction sites are amenable to being predicted by the method described in this thesis. In Tables 4.4 and 4.5, Prosite domains are identified in the predicted site for 'protein A' in the PPI. The domains are then combined in a pairwise manner with domains from the other participant protein in the PPI, protein B. Pairwise domain pairs are then aggegrated together for the whole dataset. Re-occurring domains from the predicted sites with a DM score of less than 0.20 (one of the protein sites in the PPI could have been correctly predicted) are shown, along with the percentage of sites in which it appeared (as 3 sites are selected for each PPI, and the percentage of PPIs the pair is present in).

Polypeptide chains which fit the pattern of the Src Homolog 3 (SH3) domain, which are known to be evolutionarily conserved, are common in the predictions. The function of the SH3 domain is still not well understood, despite references to it as 'Molecular Velcro' [38]. However, the interest in this domain is increased because it re-occurs often with Proline rich areas, with which it is known to interact [38]. The Src Homolog 2 (SH2) domain is also interesting in this manner as it re-occurs often with phosphorylated domains, with which it is known to interact with high affinity.

The multiple occurrences of myristoylation domains in both Tables 4.4 and 4.5 is of interest, as there are many proteins which are acylated (adding an acyl group to a protein) via a myristate [52]. If one protein in a candidate PPI catalyzes a myristoylation site, and the second protein contains a myristoylation site, there exists a good possibility for an interaction site between these two regions on the respective amino acid sequences, which would result in a PPI. Unfortunately, an exhaustive database of myristoylation catalysts does not exist, but exhaustive versions of Tables 4.4 and 4.5 could help further this research.

4.4.4 Accurate Site Predictions With Domains

While it remains of interest to identify co-occurring domains at a global scale, there are interesting results when individual confirmed and predicted sites are compared. A candidate protein is chosen (YHR016C) to demonstrate the practical applications

Freq.	% Sites	% Pairs	Domain A	Domain B
67	0.283	0.110	Casein kin. II phos.	N-myristoylation site.
63	0.278	0.097	N-myristoylation site.	Protein kin. C phos.
59	0.278	0.095	Protein kin. C phos.	Tyrosine protein kin.
49	0.231	0.062	Protein kin. domain	SH2 domain
34	0.160	0.068	Casein kin. II phos.	SH2 domain
34	0.160	0.058	Casein kin. II phos.	Proline-rich region
32	0.151	0.050	Casein kin. II phos.	Tyrosine protein kin.

Table 4.4: Most Frequently Co-occurring Domains in Lab Confirmed Interaction Sites From Human (DM Score less than 0.20)

Freq.	% Sites	% Pairs	Domain A	Domain B
72	0.626	0.163	N-myristoylation	Prot. kin. C phos.
53	0.582	0.154	Casein kin. II phos.	N-myristoylation
48	0.495	0.134	Prot. kin. C phos.	SH3 domain
36	0.396	0.126	Casein kin. II phos.	SH3 domain
33	0.341	0.089	N-glycosylation	Prot. kin. C phos.
28	0.286	0.073	N-glycosylation	Casein kin. II phos.
28	0.275	0.069	N-glycosylation	N-myristoylation
24	0.264	0.073	N-myristoylation	SH3 domain
22	0.242	0.081	N-glycosylation	SH3 domain
22	0.242	0.081	Casein kin. II phos.	Proline-rich region

Table 4.5: Most Frequently Co-occurring Domains in Lab Confirmed Interaction Sites From Yeast (DM Score less than 0.20)

Protein A	Domain A	Range A	Protein B	Domain B	Range B	DM
YHR016C	SH3	426-467	YCR088W	SH3	499-535	0.02
YHR016C	SH3	407-467	YLR144C	Glycosidases	15-36	0.01
YHR016C	SH3	400-467	YMR192W	N/A	187-214	0.02
YHR016C	SH3	407-467	YPL249C	N/A	280-301	0.02

Table 4.6: Predicted Interaction Sites and Associated Domains for YHR016C and partners

and accuracy of the prediction of interaction sites.

Table 4.6 displays the DM scores by using the PIPE matrices and the peak selection method described in 3.5.2. Also included in the table are the domains within the predicted interaction site, obtained from the Saccharomyces Genome Database (SGD) [6], and the predicted sequence ranges. All four PPIs in are nearly exact predictions (low DM scores). Figure 4.14 provides a visual depiction of a DM score of 0.02, where the predicted range slightly overlaps the lab-confirmed range.

In all the PPIs, approximately the same range is predicted for YHR016C. When compared with the lab-confirmed (411-468, [6]), they are all very near matches. The predicted range contains the Src homology 3 (SH3) domain, known to be a conserved sequence, see Figure 4.12. It is worth noting that the site predicted in YMR192W 4.14 is conserved globally (because of the high PIPE matrix score); however, it has not been identified by Prosite or by other large sequence annotation databases PFam [13], SMART [34], or PANTHER [51] as a recognized domain or motif, (see Figure 4.13). The ability of PIPE to accurately predict confirmed interacting regions not previously known to be domains or motifs is functionality not possible in other sequence based interaction site predictors.

4.5 Interaction Site Features

Several characteristics of amino acid sequences have shown to be indicative of interactions sites in Chapter 2. For a given amino acid sequence, several scales and tools exist which can be used to calculate the hydropathy, charge, solvent accessibility, and complexity. Using the subsequences identified by the method described in 3.5.2, the scales and techniques are applied, and a score for each property is calculated. Using

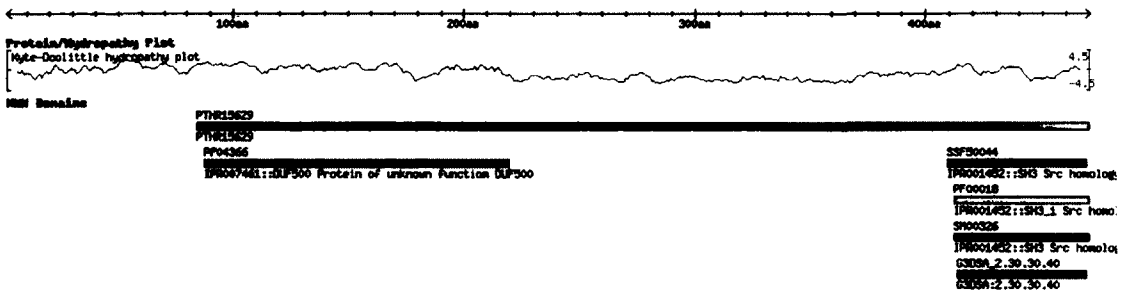

Figure 4.12: SGD info for YHR016C

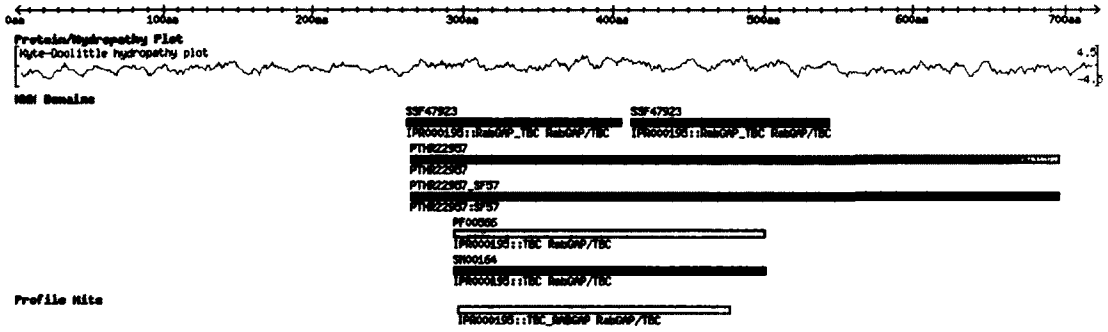

Figure 4.13: SGD info for YMR192W

the scores from each property, it may be possible to identify which peak for a given PPI would be the most likely to interact. This was tested using the four features identified below.

4.5.1 Hydropathy Feature

The scale shown in Table 4.7 is that of [31] and is used to generate the hydropathy feature score. Higher values in the scale indicate a more hydrophobic amino acid, and lower values indicate more hydrophilic amino acids. Using the chosen interacting subsequences for each member of an interacting pair, and hydropathy value for each amino acid in the subsequence, an overall score for each subsequence is calculated, and then combined together in equation 4.5.

$$HydropathyScore = \sum hydropathy_{siteA} + \sum hydropathy_{siteB} \qquad (4.5)$$

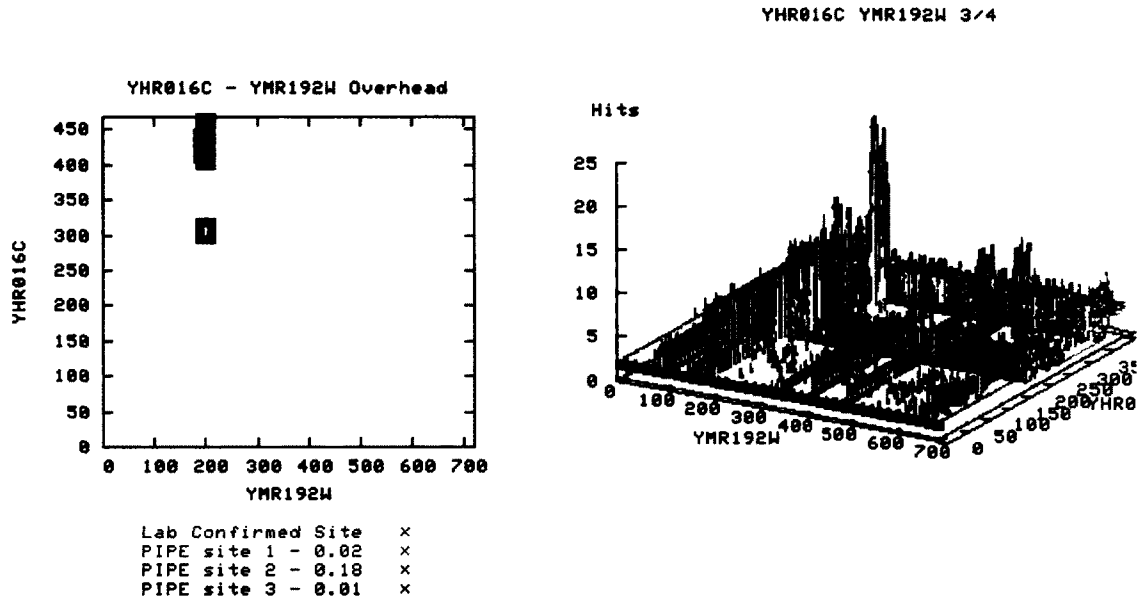

Figure 4.14: PIPE Predicted Sites for YHR016C-YMR192W (S. cerevisia)

Figures 4.15 and 4.16 displays the distribution of Hydropathy scores for both human and yeast, respectively. Negative scores indicate that the combined hydropathy of both interaction sites in the participant proteins was hydrophilic. Both the lab-confirmed sites and predicted sites for human and yeast show a significant tendency towards hydrophilic sites. This is consistent with previous research indicating that hydrophilic regions can be used as a guide for predicting interaction sites. However, when random data is introduced, no clear differentiation between the random sites,

Amino Acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Hydropathy	1.8	2.5	-3.5	-3.5	2.8	-0.4	-3.2	4.5	-3.9	3.8	1.9	-3.5	-1.6	-3.5	-4.5	-0.8	-0.7	4.2	-0.9	-1.3

Table 4.7: Kyte/Doolittle Hydropathy Scale used in Hydropathy Feature.

Amino Acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Charge	6.01	5.05	2.85	3.15	5.49	6.06	7.60	6.05	9.60	6.01	5.74	5.41	6.30	5.65	10.76	5.68	5.60	6.00	5.89	5.64

Table 4.8: Charge Scale used in Charge Feature.

and lab-confirmed sites or predicted sites is apparent. This leads to the assumption that the hydropathy score used in this thesis would not be a good indicator of interaction site activity at a proteome-scale.

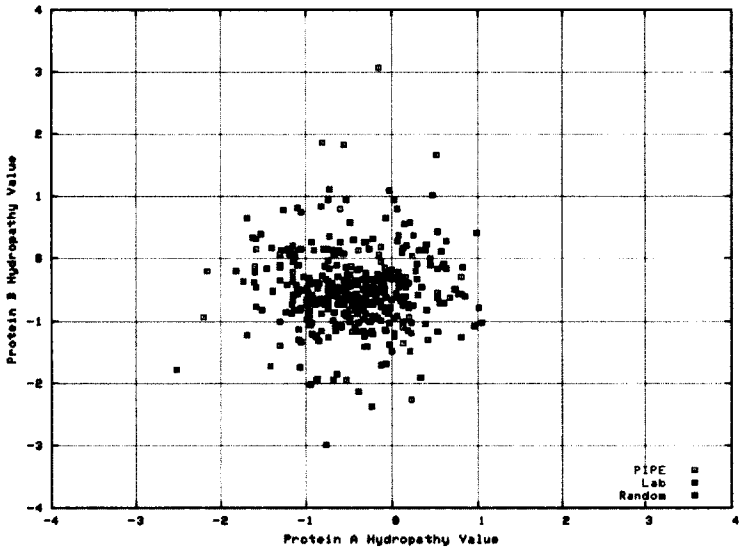

Figure 4.15: PIPE Hydropathy Scores (H. sapien)

4.5.2 Charge Feature

The standard way to measure charge is the isoelectric point (IEP) value. The IEP of an amino acid is the pH at which there is no net electric charge for the molecule. pH values below the IEP indicate positive charge, and pH values above the IEP indicate negative charge. Using the subsequence of amino acids identified by the method described in 3.5.2 and the scale defined in Table 4.8, the net charge is calculated using equation 4.6.

$$ChargeScore = \bar{charge}_{siteA} + \bar{charge}_{siteB} \tag{4.6}$$

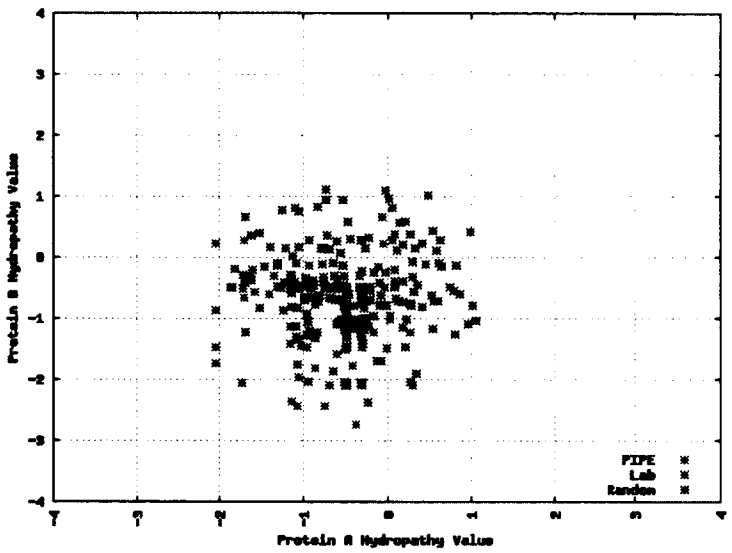

Figure 4.16: PIPE Hydropathy Scores (S. cerevisiae)

Figures 4.17 and 4.18 display the distribution of Charge scores. The lab-confirmed sites, predicted sites, and random sites for human and yeast all show a significant tendency towards scores between 5 and 7. As was the case with the Hydropathy score, when lab-confirmed and predicted scores are compared with random data, the distribution is very similar and a differentiation between random and lab-confirmed cannot be made. A similar conclusion to the Hydropathy score is reached for the Charge score: the use of the use of IEP using the charge score defined in 4.6 would not useful on proteome scale interaction site prediction.

4.5.3 Additional Features

As discussed in 2, Solvent Accessibility has been shown to have an effect on the interaction sites of PPIs. The RVP-NET algorithm [1] predicts solvent accessibility scores from sequences of amino acids using a neural network algorithm. Using the interaction sites identified in 3.5.2 as input to RVP-NET, the resulting solvent accessibilities are used in a simple solvent accessibility score defined in equation 4.7.

$$SolventAccessibilityScore = \frac{RVPNET_{siteA} + RVPNET_{siteB}}{2} \tag{4.7}$$

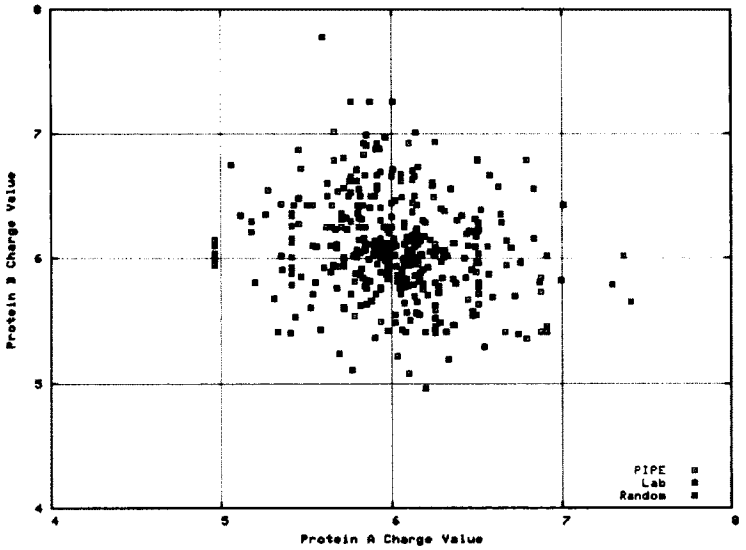

Figure 4.17: PIPE Charge Scores (H. sapien)

Sequence complexity measures the frequency and diversity of the composition of amino acids in the sequence. Using a simple algorithm developed by [55], which only uses the frequencies of the amino acids, the complexity measure of each protein is calculated. A simple complexity score is defined in equation 4.8

$$ComplexityScore = \frac{complexity_{siteA} + complexity_{siteB}}{2} \tag{4.8}$$

Neither the Complexity score or the Solvent Accessibility score exhibited any consistent pattern of scores between the predicted sites or lab-confirmed sites. In fact, among lab-confirmed sites, there was no clear distribution or separation of scores. This could be the result of the specific implementations that were chosen [55, 1], in contrast to the Hydropathy Score and Charge score which are based on standards (Tables 4.7 4.8). It could also be the basis to assume that neither feature is useful in predicting interaction sites on a proteome scale. In either case, the scores were not useful in any ranking of peaks discovered in the PIPE matrices.

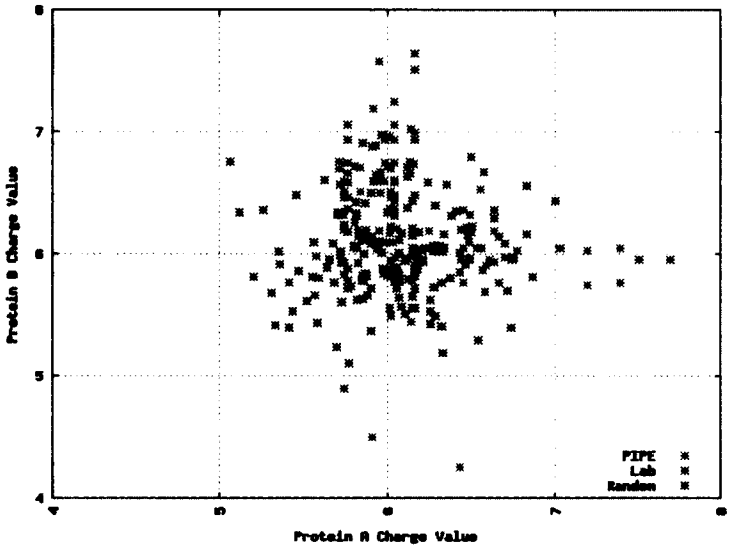

Figure 4.18: PIPE Charge Scores (S. cerevisia)

Chapter 5

Discussion of Predicted Interaction Sites

5.1 Introduction

This chapter summarizes the results obtained from performing large scale experimentation on 14438 novel yeast PPIs identified by [44], and 2347 novel human PPIs of interest identified in collaboration with researchers at the University of Leuven.

5.2 Biological Significance of Results

A dataset of 14438 novel interacting yeast proteins was obtained from [44]. In collaboration with the University of Leuven, a list of 2347 human PPIs of interest in Coronary Heart Disease (CHD) were obtained. Validation datasets from DOMINO achieved more accurate predictions when sites with peaks less than 10 were removed (see Section 4.3.2). Building on this knowledge, PPIs with PIPE matrix scores of less than 10 should not be considered in our interaction site prediction, and are removed from analysis. Of the 14438 novel yeast interactions, 5957 had peaks above 10, while 1982 of 2347 remained from the CHD dataset. Using the method described in this thesis, an interaction site would only be predicted on 41% of the novel yeast PPIs. As more PPIs are added to the PIPE database, more re-occuring sequences will be added. In turn, the percentage of PPIs having a peak higher than 10 will increase, allowing more sites to be predicted.

While it appears at first glance that predicting human interactions is done with greater coverage than yeast, it is worth considering that the novel yeast proteins were

Species	Total Pairs	Pairs with Peak > 10	Sites w/ No Prosite
S. cerevisiae	14438	5957	4417
H. sapien	2347	1982	2603

Table 5.1: Novel Protein Datasets

Freq.	% Sites	% Pairs	Domain A	Domain B
233	10	3	EGF-like domain	N-myristoylation
119	6	2	N-myristoylation	Zinc finger C2H2 domain
78	4	2	'Homeobox' domain	N-myristoylation
25	1	1	Ankyrin repeat region	EGF-like domain

Table 5.2: Predicted Co-occurring Domains From CHD Human PPIs

obtained from a proteome wide sample, whereas the human dataset was a specific, targeted family of proteins, of which the underlying PIPE PPI database had very good coverage.

5.2.1 Re-occurring Domains Within Predicted Sites

Domain information from predicted interaction sites can be used to provide insight into what interactions are amenable to being predicted by the method described in this thesis.

Domain and motif knowledge can provide information useful in evaluating the confidence of a predicted interaction site. They are also valuable to the biology community as it provides a level of confidence and more contextual knowledge as a way to explain the predicted interaction site.

While there were many similar co-occurring domains in Table 5.2 and Table 4.4, there are several which are worth noting. First, the identification of the Epidermal Growth Factor (EGF) Domain (Prosite ID: PS01186) is of interest. The EGF domain is an evolutionary conserved domain conserved over animal proteins between thirty to forty amino acids long [10, 5], and is known to bind with high affinity to cell-surface regions. The EGF domain's ability to bind with high affinity to cell-surface regions and its appearance in over 12% of all interactions in the exhaustive list warrant more investigation into this observation in order to determine its significance.

Second, the Zinc finger C2H2 domains are nucleic acid-binding protein structures ([12], [41]), however are not known to be evolutionarily conserved. Both the human proteome and the human PIN have grown considerably since the writing of these articles. With this, the conclusion reached on Zinc finger domain conservation, as well as other motifs and domains, may be changed. The Homeobox domain is a known

Freq.	% Sites	% Pairs	Domain A	Domain B
642	4	9	ATP/GTP-bind site	Casein kinase II phosphoryl.
521	452	346	ATP/GTP-bind site	Protein kinase C phosphoryl.
44	44	42	ATP/GTP-bind site	AAA-protein family
28	28	28	ATP/GTP-bind site	Microbodies C-term target signal
27	27	27	ATP/GTP-bind site	Amidation site
21	21	17	ATP/GTP-bind site	Trp-Asp (WD) repeats
18	18	17	ATP/GTP-bind site	Asparagine-rich region
11	11	11	ATP/GTP-bind site	Cell attachment sequence
9	9	9	ATP/GTP-bind site	Trp-Asp (WD) repeats
8	8	8	ATP/GTP-bind site	Glutamic acid-rich region
8	8	8	ATP/GTP-bind site	GTP-binding nuclear protein ran
4	4	4	ATP/GTP-bind site	EF-hand calcium-binding domain

Table 5.3: Predicted Co-occurring Domains From Novel Yeast PPIs

well-conserved domain responsible for DNA binding [17]. The domain is relatively large at 60 amino acids and noteworthy that such a large domain can be identified via the PIPE method; however, it is not known what partners it prefers to bind with.

Finally, Ankyrin repeat region profiles are known to function as protein-protein interactions [4], and are of interest in this context they also appear in large variety of functionally diverse proteins. An exhaustive list of co-occurring domains in the CHD dataset can be found in Appendix A.

Table 5.3 highlights some of the interesting co-occurring domains. The CHD human dataset revealed a proportionally much wider variety of domains, 82 domains in 1982 PPIs, when compared to the novel yeast dataset, 143 domains in 5957 PPIs. This could be due to a larger variety of domains in human proteins or the selective choosing of candidate PPIs in the CHD dataset. While the breadth of domains is not impressive with the yeast dataset , it is of interest that the ATP/GTP domain (Prosite ID: PS00017) appeared in over 14% of the interactions and is a known binding domain. The ATP/GTP domain co-occurs with 36 other domains, some of which are shown in 5.3, and are all candidates for an acurate binding site prediction. An exhaustive list of co-occurring domains in the novel yeast dataset can be found in Appendix A.

Species	Peak Height	Sequence
S. cerevisiae	427	ILDGDEDEPEEEDENEGDDEEDTYDS
S. cerevisiae	285	DADGDDQTEEGEVEKEQKEEDEEEGPK
S. cerevisiae	266	AFDNDESDAQDDANNEKEDDGEEF
S. cerevisiae	215	ADQDVEGEDEGGDAIENEDEDEDPSPS
H. sapien	504	GQVTPPTPPQTAQPPLPGPPPAAVE
H. sapien	320	IQNLERGYRMVRPDNCPEELYQLMRLCWKERPEDR
H. sapien	234	AAIEPQPSPPHSEPPSVEQPPKPK
H. sapien	195	PIWLQPSPPPQSSPPPQPHP

Table 5.4: Non-Prosite Re-occurring Sequences From Predicted Interaction Sites

5.2.2 Novel Motifs

Following peak identification (using 3 peaks), interaction site subsequences were cross-referenced in Prosite for similarity to existing domains and motifs. The number of interaction site subsequences not containing an identified domain or motif in Prosite totaled 2603 and 4417 for human and yeast, respectively. An exhaustive list can be found in Appendix A. Table 5.4 displays some of the highest peaks (most re-occurrent sequences) not contained in Prosite from both the yeast and human datasets.

5.3 Comparison With Other Techniques

Proteome-scale interaction site prediction has not yet been heavily researched or implemented. To the best of knowledge, no other interaction site prediction method exists at the proteome-scale, using individual proteins, with interaction sites predicted on both proteins. Thus it is very difficult to make direct, empirical comparisons, as methods surveyed in Chapter 2 are all missing one of the characteristics above.

The differences among current interaction site prediction methods and the one described in this thesis are numerous (the scale at which they operate notwithstanding). All studies surveyed couple the prediction of PPI with the prediction of an interaction site [29, 45, 54, 36, 39, 21, 20, 35, 56, 40, 19, 47, 33, 24, 27, 49]. In the sensitivity and specificity calculations, the site is assumed correct if the PPI is correctly predicted. Conversely, if no interaction site is found, then a PPI is thought not to occur. This differs from the approach taken in PIPE, where the PPI prediction and site prediction are de-coupled. PIPE can predict a PPI without having a predicted interaction site

(for instance if there is no peak with a PIPE score above 10, a site is not predicted). This is significant as the specificity and sensitivity scores for PIPE are not affected by the configuration and selection of interaction sites.

A gold standard or training set of interaction site characteristics, domains, or motifs is required for prediction of interaction sites. The prediction of interaction sites is based entirely on the observed interaction site characteristics of the training data. This approach may work for similar families of proteins or interactions within a small group of proteins for identifying the large complexes; however, if a protein does not fit a profile of the training, results will be spurious. This contrasts with the PIPE method, which can identify interacting domains, as well as novel sub-sequences not previously reported as domains or motifs, without the need for interaction site training data.

The granularity at which current prediction techniques operate is typically at the family or complex level. Only one study made predictions at the individual protein level [54]. Even these predictions were constrained in that only the site of one participant proteins in the PPI was predicted (Motif M of Protein A binds to Protein B). Additionally, if multiple instances of the same motifs (M_1, M_2) occur on a sequence, no differentiation is made, and ambiguity between which motif is actually responsible for the interaction remains. The PIPE method readily predicts the interaction site between any two proteins, regardless of family, and provides a predicted interaction site on both proteins. PIPE also incorporates contextual knowledge of a motif's surrounding amino acids to differentiate between multiple motifs on a single protein.

Chapter 6

Summary of work, Conclusions and Future Work

6.1 Introduction

This chapter lists the conclusions and major contributions of this thesis, as well as future work to be done.

6.2 Conclusions

- Interaction sites can be predicted from re-occurring polypeptide sequences. While the method performed significantly better than random sites in both validation datasets, the yeast dataset performed better than the human. Both human and yeast PIPE databases were in the range of the same size, approximately 40,000 and 36,000 confirmed PPIs respectively, but the difference was likely due to more extensive and wider range of PPIs making up the yeast database. The lack in quantity of lab-confirmed and predicted interaction sites only underscores the value of an accurate proteome scale interaction site prediction method.

- Potential novel motifs can be identified from re-occurring polypeptide sequences. As discussed in Section 4.4.4, there are domains and motifs which have not yet been identified as globally significant, but can have a role in determining interaction sites. The validation of new motifs and domains is an extensive wet lab process which requires significant time, resources, and willing collaborators.

- Using equations defined in 4.5 and 4.6, no significant difference was evident between lab, predicted, and random interaction sites for the hydropathy and charge property scores. The scores can not be used to more effectively rank the predicted interaction sites. Solvent accessibility and complexity scores did not

yield any consistent pattern within lab sites and thus predicted values do not have an effective comparator to separate good predictions from bad predictions.

6.3 Contributions

- Developed an algorithm (see Section 3.5.2) to identify potential interaction sites and novel motifs, from re-occurring polypeptide sequences. To the best of knowledge, there is no other interaction site prediction method which can predict interaction sites on a proteome wide scale. Thus, there have been 7939 predicted interaction sites on novel PPIs from both human and yeast proteins. The discovery of 7020 over-represented sequences in both human and yeast PPI datasets will require the willing collaboration of researchers with wet lab resources to validate their usefulness and function.

- Developed two empirical metrics (see Sections 4.3.1 and 4.3.3) . The Size Measure (SM) is used to evaluate whether the interaction sites being predicted are of a similar size as the validation data. This allows the configuration of the basic peak selection algorithm to select interaction sites of a size consistent with the validation data. The Distance Measure (DM) is used to evaluate a predicted interaction site's proximity to a lab-confirmed interaction site. It is envisioned that future interaction site prediction engines can use this metric for comparative analysis with other techniques. To the best of knowledge, no method exists to evaluate or compare interaction site prediction techniques at the individual protein level.

- Identified a set of Prosite domains which characterizes the known interacting domains that PIPE can identify. The domains are used to provide biological context and significance to interaction sites in human and yeast PPIs.

- Identified a set of potential novel motifs in human (2603) and yeast (4417) proteins. These potential motifs are polypeptide sequences predicted as an interaction sites, meaning they are globally significant re-occurring sequences,

but do not match any pattern in the Prosite database. Validation of these sequences require extensive wet lab resources and willing collaborators.

6.4 Future Work

Future direction and work can be done to further refine the selection, identification, and filtering of interaction sites. Additional features could be added to form a ranking of predicted sites, and lead to the exclusion of inaccurate predictions.

- From the primary structure of a protein, the secondary and tertiary structures can be determined. Structural motifs such as beta-sheets which are known to be favored by interaction sites among proteins [57] can be discovered by solving the secondary structure of a protein. Using the methods used to solve the 3D structure of a protein, predicted interaction sites could be folded and analyzed for the existence of known structural motifs and the existence of novel ones. The identification of structural motifs within predicted interaction sites could yield more accurate predictions and identify re-occurring secondary structures.

- Within the primary structure of a protein, there exists low-complexity regions or sub-sequences. These low-complexity regions contain a non-diverse amino acid composition and are not thought to be functionally significant. However, most protein sequences contain some complexity regions, and because of their non-diverse nature, larger low-complexity regions can appear as globally recognized re-occurrent sequences and result in peaks in a PIPE matrix. Using the peak selection algorithm in this thesis, the selection of a peak which represents a low-complexity region will yield an inaccurate prediction. A complexity filter could be used to mitigate the selection of low complexity regions, and filter out unlikely and inaccurate predictions. The use of a complexity filter could also improve the overall specificity of PIPE by eliminating the highest occurring low-complexity peak and basing the PPI prediction on the new scores.

- In addition to the above features, there are many more additional features which could be added. Molecular weight [7] and solvent accessibility [1], to name a

few, have shown to be use useful in predicting interaction sites, and should be considered for addition. With the addition of so many features, it becomes more complex to classify interaction sites as accurate. Classification is a well studied problem, and the use of Support Vector Machines (SVMs) has been used for interaction site prediction in the past ([29],[37],[39],[45]) and can be leveraged to provide a more sophisticated approach to identifying meaningful and useful features.

- Using Prosite for domain pattern recognition on the interaction site sequence has yielded some interesting biological properties. Domains such as SH3, SH2, and myristoylation can provide context to a PPI prediction. Incorporating domain information into the PIPE DB as annotations during the window comparison could potentially improve prediction accuracy of PPIs by providing a level of confidence.

Appendix A

Electronic Appendix

In order to make datasets and code more easily accessible, an electronic appendix is hosted at http://www.dehne.carleton.ca/Members/adam/thesis-appendix/

Bibliography

[1] S. Ahmad, M. M. Gromiha, and A. Sarai. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, 19(14):1849–1851, 2003.

[2] P. Aloy and et al. Protein complexes: structure prediction challenges for the 21st century. *Struct. Biol.*, 15:15–22, 2005.

[3] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5896–5901, 2002.

[4] P. Bork. Hundreds of ankyrin-like repeats in functionally diverse proteins: Mobile modules that cross phyla horizontally? *Proteins: Structure, Function, and Genetics*, 17(4):363–374, 1993.

[5] P. Bork, A. K. Downing, B. Kieffer, and I. D. Campbell. Structure and distribution of modules in extracellular proteins. *Quarterly Reviews of Biophysics*, 29(02):119–167, 1996.

[6] J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer, and D. Botstein. Genetic and physical maps of saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73, 1997.

[7] K.-i. Cho, D. Kim, and D. Lee. A feature-based approach to modeling protein-protein interaction hot spots. *Nucl. Acids Res.*, 37(8):2672–2687, 2009.

[8] L. Conte and et al. The atomic structure of proteinprotein recognition sites. *J. Mol. Biol.*, 285(285):21772198, 1999.

[9] E. de Castro, C. J. A. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch, and N. Hulo. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucl. Acids Res.*, 34(suppl 2):W362–365, 2006.

[10] A. K. Downing, V. Knott, J. M. Werner, C. M. Cardy, I. D. Campbell, and P. A. Handford. Solution structure of a pair of calcium-binding epidermal growth factor-like domains: Implications for the marfan syndrome and other genetic disorders. *Cell*, 85(4):597 – 605, 1996.

[11] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299:371–374, 1982/09/23/print.

[12] R. M. Evans and S. M. Hollenbergt. Zinc fingers: Gilt by association. *Cell*, 52(1):1 - 3, 1988.

[13] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucl. Acids Res.*, 36(suppl1):D281–288, 2008.

[14] X. Gallet, B. Charloteaux, A. Thomas, and R. Brasseur. A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology*, 302(4):917 - 926, 2000.

[15] X. Gallet and et al. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, 302:917926, 2000.

[16] A.-C. Gavin and et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[17] W. J. Gehring and Y. Hiromi. Homeotic genes and the homeobox. *Annual Review of Genetics*, 20(1):147–173, 1986.

[18] M. Guharoy and P. Chakrabarti. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15447–15452, 2005.

[19] K. Guimaraes, R. Jothi, E. Zotenko, and T. Przytycka. Predicting domain-domain interactions using a parsimony approach. *Genome Biology*, 7(11):R104, 2006.

[20] J. Guo, X. Wu, D.-Y. Zhang, and K. Lin. Genome-wide inference of protein interaction sites: lessons from the yeast high-quality negative protein-protein interaction dataset. *Nucl. Acids Res.*, 36(6):2002–2011, 2008.

[21] Y. S. Huan-Xiang Zhou. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Genetics*, 44(3):336–343, 2001.

[22] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist. The prosite database. *Nucl. Acids Res.*, 34(suppl 1):D227–230, 2006.

[23] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 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):4569–4574, 2001.

[24] Z. Itzhaki, E. Akiva, Y. Altuvia, and H. Margalit. Evolutionary conservation of domain-domain interactions. *Genome Biology*, 7(12):R125, 2006.

[25] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. E. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. Capri: A critical assessment of predicted interactions. *Proteins: Structure, Function, and Genetics*, 52(1):2–9, 2003.

[26] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformation of amino acid side-chains in proteins. *Journal of Molecular Biology*, 125(3):357 – 386, 1978.

[27] R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of Molecular Biology*, 362(4):861 – 875, 2006.

[28] M. G. Kann. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, page bbm031, 2007.

[29] A. Koike and T. Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering, Design and Selection*, 17(2):165–173, 2004.

[30] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrn-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[31] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105 – 132, 1982.

[32] S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande. Folding@home and genome@home: Using distributed computing to tackle previously intractable problems in computational biology. *Computational Genomics, Horizon Press*, 2002.

[33] H. Lee, M. Deng, F. Sun, and T. Chen. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 7(1):269, 2006.

[34] I. Letunic, T. Doerks, and P. Bork. SMART 6: recent updates and new developments. *Nucl. Acids Res.*, 37(suppl1):D229–232, 2009.

[35] H. Li, J. Li, and L. Wong. Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.

[36] M.-H. Li, L. Lin, X.-L. Wang, and T. Liu. Protein protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5):597–604, 2007.

[37] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking of protein residues by importance. *J. Mol. Bio.*, 5(336):1265–1282, 2004.

[38] C. J. Morton and I. D. Campbell. Sh3 domains: Molecular 'velcro'. *Current Biology*, 4(7):615 – 617, 1994.

[39] Y. Ofran and B. Rost. Predicted protein protein interaction sites from local sequence information. *FEBS letters*, 544(1):236–239, 2003.

[40] Y. Ofran and B. Rost. ISIS: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–16, 2007.

[41] F. Payre and A. Vincent. Finger proteins and dna-specific recognition: Distinct patterns of conserved amino acids suggest different evolutionary modes. *FEBS Letters*, 234(2):245 – 250, 1988.

[42] S. Pitre, M. Alamgir, J. Green, M. Dumontier, F. Dehne, and A. Golshani. Computational methods for predicting proteinprotein interactions. *Advances in Biochemical Engineering/Biotechnology: Protein Protein Interaction*, 110:247–267, 2008.

[43] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J.Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, 7:365 (15 pages), 2006, available via PubMed at http://www.biomedcentral.com/pubmed/16872538.

[44] S. Pitre, C. North, M. Alamgir, M. Jessulat, A. Chan, X. Luo, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani. Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucl. Acids Res.*, 36(13):4286–4294, 2008.

[45] I. Res, I. Mihalek, and O. Lichtarge. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10):2496–2501, 2005.

[46] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotech*, 17(10):1030–1032, 1999.

[47] R. Riley, C. Lee, C. Sabatti, and D. Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, 6(10):R89, 2005.

[48] D. W. Ritchie. Recent progress and future directions in protein-protein docking. *Current Protein and Peptide Science*, 9(1):1–15, 2008.

[49] S.-E. Schelhorn, T. Lengauer, and M. Albrecht. An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16):i35–41, 2008.

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

[51] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, 13(9):2129–2141, 2003.

[52] D. A. Towler, J. I. Gordon, S. P. Adams, and L. Glaser. The biology and enzymology of eukaryotic protein acylation. *Annual Review of Biochemistry*, 57(1):69–97, 1988.

[53] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

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

[55] J. C. Wootton. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Computers and Chemistry*, 18(3):269 – 285, 1994.

[56] C. Yan, D. Dobbs, and V. Honavar. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20(suppl):i371–378, 2004.

[57] H.-X. Zhou and S. Qin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209, 2007.