USING TEXT MINING OF PUBMED ABSTRACTS AS AN EVIDENCE SOURCE IN COMPUTATIONAL PREDICTION OF WW DOMAIN-MEDIATED PROTEIN-PROTEIN INTERACTIONS

MARINA OLHOVSKY

A Thesis Submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

Graduate Program in Biology

York University, Toronto, Ontario

AUGUST 2015

ABSTRACT

Protein-protein interactions (PPIs) are a key regulatory mechanism in coordinating a multitude of processes vital to normal cellular function. There exist a number of wet-lab small-scale and high-throughput methods for accurately identifying PPIs; however, despite their accuracy, these methods are expensive both in terms of time and finances. Complementing experimental methods with computational predictions increases the effectiveness of wet-lab small scale methodologies in identifying high quality protein interaction networks. Computational predictions are made by applying bioinformatics and machine-learning algorithms to large-scale training sets obtained from wet-lab experiments, or by extracting information on PPIs from high volumes of published data that do not directly identify protein interactions but are nonetheless correlated with them. A disadvantage of computational predictions is their high degree of inaccuracy, namely too many false positives and false negatives. To improve the accuracy of computational predictions, it is important to consider interactions that are likely to occur *in vivo* under certain biological conditions, termed “context”. One technique for improving prediction accuracy is analyzing data obtained via different types of experiments that consider different features of the co-occurring proteins, such as co-localization, co-expression, correlated mutations, or semantic similarity. These experimental sources and their resulting data are called “sources of evidence”. Integrating data from multiple independent supporting evidence sources improves prediction accuracy.

In this work, I used text mining of PubMed abstracts as an evidence source for protein interactions. I hypothesized that proteins whose names are frequently mentioned in the same abstract are more likely to interact *in vivo* compared to randomly chosen proteins. A comparison of three text mining techniques – gene name co-occurrence, MeSH term indexing, and co-occurrence with a controlled vocabulary – shows that co-occurrence with a controlled vocabulary yields the highest precision and recall. I concluded that gene name co-occurrence with a controlled vocabulary can, therefore, be used as a novel evidence source for prediction of WW domain-mediated PPIs.

DEDICATION

This M. Sc. work is dedicated to the eternal memory of my mother, **LARISA OLHOVSKY**, who is no longer with us but will never be forgotten.

ACKNOWLEGMENTS

This Masters work is dedicated, first and foremost, to the eternal memory of my mother, **LARISA OLHOVSKY**, who did everything in human power and beyond to help me get to where I am today. She is no longer on this earth, but her memory has been my guiding light every step of the way for the past 10 years. She valued education above all, and my graduate degree is one tribute to her. She passed away from triple-negative breast cancer; my professional and academic efforts since her passing have been focused on helping bring scientists one step closer to conquering this devastating disease.

When Mom was alive, I joined the laboratory of Dr. **TONY PAWSON**, a distinguished scientist, taken from us too fast too soon. Working in Tony's lab has been an extreme privilege, a valuable learning experience that has given me a springboard into my dream of working in the field of bioinformatics and computational biology. Tony's genius and contribution to my success is never forgotten.

I wholeheartedly thank Dr. **KAREN COLWILL**, a staff scientist in the Pawson lab, who has been a mother-like figure for me these past 10 years, since Mom's passing. My life during that time has been centered around work at Mt. Sinai hospital, and Karen played a fundamental role in my professional and personal success.

I express sincere gratitude to the members of my supervisory committee. My greatest thanks go to Dr. **GARY BADER**, my co-supervisor from the University of Toronto, who has assumed the role of an adjunct professor at York University for the supervision of my Master's. Gary has devotedly led me to defense, with our regular meetings, his travelling to York U from downtown Toronto for our yearly committee meetings, e-mail consultations, advice and verification of work.

I want to thank Dr. **LOGAN DONALDSON** for all his support throughout the years, from the Proteomics course to my Master's defense.

A big thank you goes out to Dr. **RONALD PEARLMAN**, a distinguished researcher at York University, who has permitted me to do my thesis under Gary's co-supervision, and has provided me with valuable guidance throughout my academic years.

Finally, my most sincere, deepest gratitude goes out to my father, **YURI OLHOVSKY** – for not letting me lose my zest for life and knowledge after Mom's passing; for being a pillar I could hold on to during any storm at these times; for believing in me, supporting me all through the ups and downs of the past 10 years; for maintaining my faith in my own abilities, for raising my spirits when I needed a lift, and for helping me get to this point today. I love you more than anyone in this world!

Special thanks to Dr. **MATTHEW H. HERYNK** for critical reading of the thesis.

TABLE OF CONTENTS

ABSTRACT .... **ii**

DEDICATION .... **iv**

ACKNOWLEGMENTS .... **v**

LIST OF ABBREVIATIONS .... **x**

LIST OF TABLES .... **xii**

LIST OF FIGURES .... **xiii**

CHAPTER 1. INTRODUCTION **1**

1.1 – Protein-protein interactions in signaling networks **1**
1.2 – Peptide Recognition Modules **3**
Figure 1.2.1: SH2 Domain 4
Figure 1.2.2: SH3 Domain 5
Figure 1.2.3: PDZ Domain 6
Figure 1.2.4: PTB Domain 7
1.3 – WW Domains **8**
Figure 1.3.1: WW Domain 9
1.4 – Experimental PPI identification **10**
Figure 1.4.1: Yeast Two-Hybrid (Y2H) PPI detection method 11
Figure 1.4.2: Co-immunoprecipitation (Co-IP) PPI detection method 13
1.5 – Computational PPI Predictions **14**
1.5.1 – The need for computational predictions 14
1.5.2 – PPI prediction methods 14
Figure 1.5.1: Position-weight matrix and sequence logo 16
1.7 – The downside of computational predictions **17**
1.6 – Sources of evidence for PPIs **17**
Figure 1.6.1: STRING database 18
Figure 1.6.2: The PIPs repository 19
1.7 – Text Mining **20**
1.8 – Precision and Recall **21**
1.9 – Thesis Outline **21**

CHAPTER 2. GENE NAME CO-OCCURRENCE **23**

2.1 – MATERIALS AND METHODS **24**
2.2 – PROCEDURE **25**
2.2.1. Datasets: 25
2.2.2. Text mining: 29
2.3 – RESULTS **30**
Figure 2.3.1: Example of protein name co-occurrence in abstracts. 30
Figure 2.3.2: A diagram illustrating the interaction between Smurf1, Smurf2, Smad1, Smad2 and Smad7 proteins from the GeneMANIA prediction server (Warde-Farley et al. 2010) (http://www.genemania.org) 31
Figure 2.3.3: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Zhang et al. 2001) 32
Figure 2.3.4: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Fukasawa et al. 2004). 33

Figure 2.3.5: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin et al. 2000) 34

Table 2.3.6: WW domain-containing protein interactions, grouped by the number of abstracts cutoff. 36

Table 2.3.7: WW domain-containing protein interaction prediction results based on co-occurrence of protein names in PubMed abstracts (2012) 37

Figure 2.3.8: Precision-recall graph of WW domain-containing PPI predictions by text mining based on co-occurrence of protein names in PubMed abstracts 38

2.4 – DISCUSSION 39

2.5 – CONCLUSIONS 39

CHAPTER 3: MeSH TERM INDEXING 40

3.1 – MATERIALS AND METHODS 41

3.1.1. Datasets: 41

3.1.2. Text mining: 43

Table 3.1.1: A sample MySQL table storing mapping between HGNC symbols and their corresponding MeSH terms: 44

Table 3.1.2: A sample MySQL table storing predicted interactions for MeSH terms indexed in the same abstract 45

Table 3.1.3: WW domain-containing protein interaction prediction results based on MeSH term indexing in PubMed abstracts (2013) 45

Table 3.1.4: Precision and recall computed for results in Fig. 3.1.3 45

Figure 3.1.5: Precision-recall graph of WW domain-containing PPI predictions by text mining based on MeSH terms indexed in PubMed abstracts 47

Table 3.1.6: Comparison of text mining prediction results based on gene name co-occurrence in abstracts and results based on MeSH term indexing 48

Figure 3.1.7: Comparison of text mining results based on gene name co-occurrence to results based on MeSH term indexing. 49

3.1.3 – DISCUSSION 50

3.1.4 – CONCLUSIONS 50

CHAPTER 4. CO-OCCURRENCE WITH CONTROLLED VOCABULARY 51

4.1. MATERIALS AND METHODS 52

4.1.1. Datasets: 52

4.1.2. Text mining: 53

Figure 4.1.1: An abstract that contains controlled vocabulary term 'PHOSPHORYLATION'. Co-occurring protein names HSF1 and BAG3 are highlighted. These proteins were predicted to interact 54

Figure 4.1.2: A Cytoscape (Shannon et al. 2003) diagram illustrating the interaction between BAG3 and HSF1 proteins 55

Table 4.1.3: The number of TPs and FPs for each individual CV term 56

Figure 4.1.4: Precision-recall table and graph for CV term 'protein interaction' 57

Figure 4.1.5: Precision-recall table and graph for CV term 'WW domain' 58

Figure 4.1.6: Precision-recall table and graph for CV term 'PPXY' 59

Figure 4.1.7: Precision and recall values for all CV terms combined 60

Figure 4.1.8: Precision-recall graph for gene name co-occurrence with all CV terms combined 61

Figure 4.1.9: Comparison of precision and recall values for gene name co-occurrence with and without all CV terms combined 62

Figure 4.1.10: Comparison of all text mining results – based on gene name co-occurrence without CV, based on MeSH terms, and based on gene name co-occurrence with CV 63

4.2. DISCUSSION 64

4.3. CONCLUSIONS 64

FUTURE DIRECTIONS 65

Figure 5.1: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin et al. 2000) 66

BIBLIOGRAPHY 67

Appendix A – Database ER diagram 72

Appendix B – Ensembl repository search results, using the Biomart query engine: 73

Appendix C – Ensembl repository search results for WW domain 74

Appendix D – PubMed abstracts in XML format 75

Appendix E – A Python script to parse downloaded abstract search results in XML format and store in the database 76

Appendix F – iRefIndex PPI dataset, used as the benchmark set in PPI prediction script (ww_benchmark.tsv) 77

Appendix G – A Python script to detect co-occurrence between WW domain-containing proteins and other human proteins in PubMed abstracts 78

Appendix H – A Python script to compare predictions to benchmark and detect TPs and FPs 83

Appendix I – Python script to retrieve MeSH terms for every human protein-coding gene. 85

Appendix J – Python script for retrieving all MeSH abstracts in MEDLINE format and identifying the MeSH terms indexed in them as Supplementary Concepts. 88

LIST OF ABBREVIATIONS

AA	–	Amino Acid
Ab	–	Antibody
AD	–	Activation Domain
BD	–	Binding Domain
Co-IP	–	Co-Immunoprecipitation
CV	–	Controlled Vocabulary
DBD	–	DNA Binding Domain
DBMS	–	Database Management System
ER	–	Entity-relation (diagram)
FN	–	False Negatives
FP	–	False Positives
HGNC	–	Human Gene Nomenclature Committee
HMM	–	Hidden Markov Model(s)
HUPO	–	Human Proteome Organization
IP	–	Immunoprecipitation
ME	–	Maximum Entropy
MeSH	–	Medical Subject Heading
NCBI	–	National Center for Biotechnology Information
NER	–	Named Entity Recognition
NIH	–	National Institutes of Health

NLM	–	National Library of Medicine
NLP	–	Natural Language Processing
OS	–	Operating System
PPI	–	Protein-protein interaction(s)
PSI	–	Proteomics Standard Initiative
PWM	–	Position-Weight Matrix
pY	–	PhosphoTyrosine
SH2	–	Src Homology 2 (domain)
SH3	–	Src Homology 3 (domain)
STAT	–	Signal Transducer and Activator of Transcription (protein)
STRING	–	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	–	Support Vector Machine(s)
TF	–	Transcription Factor(s)
TP	–	True Positive(s)
Y2H	–	Yeast Two-Hybrid

LIST OF TABLES

Table 2.3.6: WW domain-containing protein interactions, grouped by the number of abstracts cutoff

Table 2.3.7: WW domain-containing protein interaction prediction results based on co-occurrence of protein names in PubMed abstracts (2012)

Table 3.1.1: A sample MySQL table storing mapping between HGNC symbols and their corresponding MeSH terms

Table 3.1.2: A sample MySQL table storing predicted interactions for MeSH terms indexed in the same abstract

Table 3.1.1: A sample MySQL table storing mapping between HGNC symbols and their corresponding MeSH terms

Table 3.1.2: A sample MySQL table storing predicted interactions for MeSH terms indexed in the same abstract.

Table 3.1.3: WW domain-containing protein interaction prediction results based on MeSH term indexing in PubMed abstracts (2013)

Table 3.1.4: Precision and recall computed for results in Fig. 3.1.3

Table 3.1.6: Comparison of text mining prediction results based on gene name co-occurrence in abstracts and results based on MeSH term indexing

Table 4.1.3: The number of TPs and FPs for each individual CV term

LIST OF FIGURES

Figure 1.1.1: Cellular signal transduction involving STATs that affects gene regulation

Figure 1.2.1: SH2 Domain

Figure 1.2.2: SH3 Domain

Figure 1.2.3: PDZ Domain

Figure 1.2.4: PTB Domain

Figure 1.3.1: WW Domain

Figure 1.4.1: The Yeast Two-Hybrid (Y2H) PPI detection method

Figure 1.4.2: Co-immunoprecipitatin (Co-IP) PPI detection method

Figure 1.5.1: Position-weight matrix and sequence logo

Figure 1.6.1: STRING repository

Figure 1.6.2: The PIPs database

Figure 2.1.1: Prediction database schema

Figure 2.3.1: Examples of protein name co-occurrence in abstracts

Figure 2.3.2. A diagram illustrating the interaction between Smurf1, Smurf2, Smad1, Smad2 and Smad7 proteins from the GeneMANIA prediction server (Warde-Farley D. et al. 2010) (http://www.genemania.org)

Figure 2.3.3. An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Zhang Y et al. 2001).

Figure 2.3.4. An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Fukasawa H et al. 2004)

Figure 2.3.5. An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin X et al. 2000)

Figure 2.3.8: Precision-recall graph of WW domain-containing PPI predictions by text mining based on co-occurrence of protein names in PubMed abstracts

Figure 3.1.5: Precision-recall graph of WW domain-containing PPI predictions by text mining based on MeSH terms indexed in PubMed abstracts

Figure 3.1.7: Comparison of text mining results based on gene name co-occurrence to results based on MeSH term indexing.

Figure 4.1.1: An abstract that contains controlled vocabulary term 'PHOSPHORYLATION'. Co-occurring protein names HSF1 and BAG3 are highlighted. These proteins were predicted to interact.

Figure 4.1.2: A Cytoscape diagram illustrating the interaction between BAG3 and HSF1 proteins

Figure 4.1.4: Precision-recall table and graph for CV term 'protein interaction'

Figure 4.1.5: Precision-recall table and graph for CV term 'WW domain'

Figure 4.1.6: Precision-recall table and graph for CV term 'PPXY'

Figure 4.1.7: Precision and recall values for all CV terms combined

Figure 4.1.8: Precision-recall graph for gene name co-occurrence with all CV terms combined

Figure 4.1.9: Comparison of all text mining results – based on gene name co-occurrence without CV, based on MeSH terms, and based on gene name co-occurrence with CV

Figure 5.1: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin X et al. 2000)

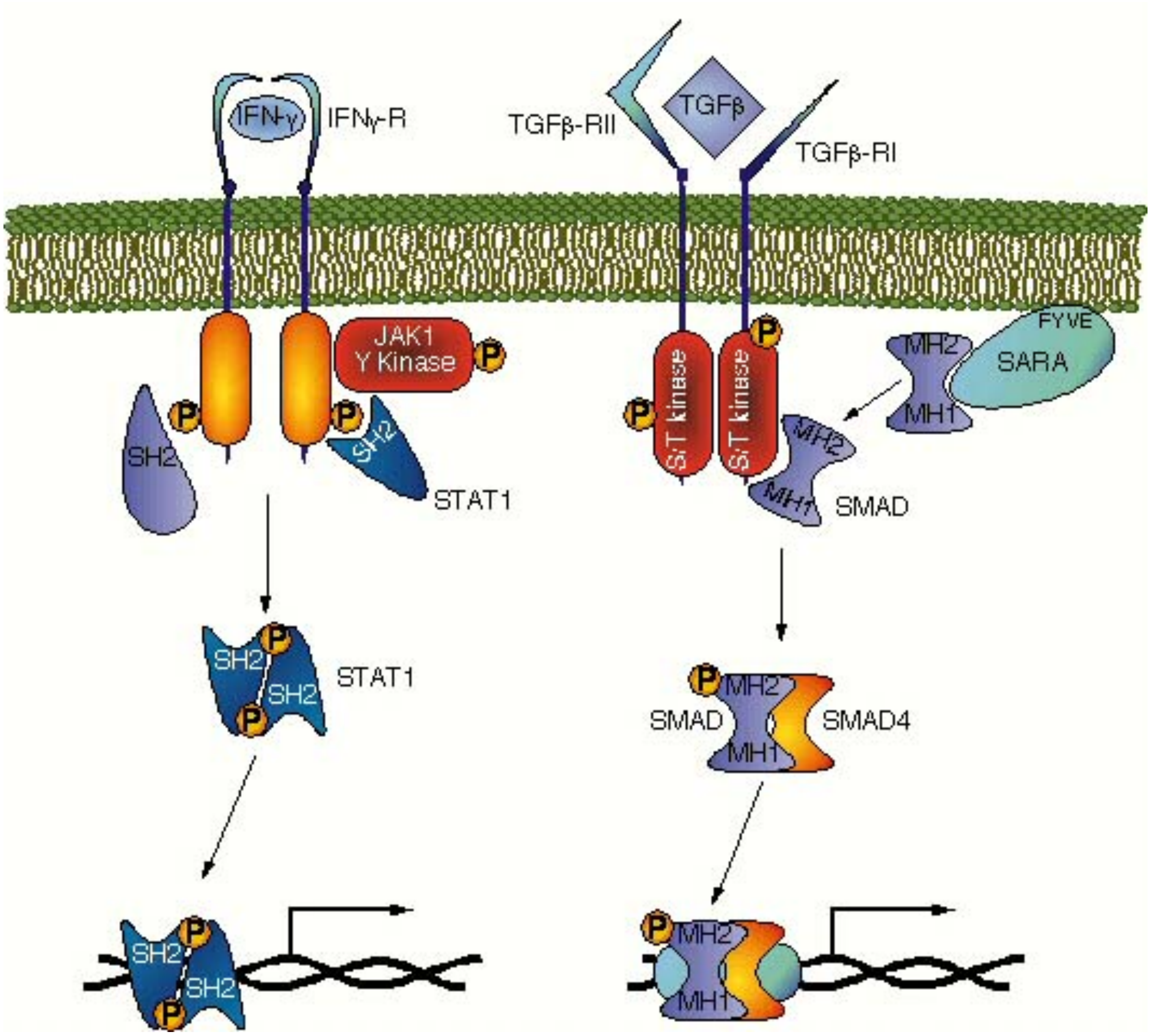
CHAPTER 1. INTRODUCTION

1.1 – Protein-protein interactions in signaling networks

In the post-genomic era, genes and their products are no longer studied only as individual entities but in the context of larger functional interaction networks within the cell. It has become increasingly evident that over 80% of gene products do not function in isolation, but form part of a coordinated cell regulatory network (Berggard T et al. 2007). An example of a cellular regulatory network is the signal transduction process [**Fig. 1.1.1**], where external stimulation events are converted into intracellular response via a series of protein interactions within the cell (Wilks and Harpur 1996). Extracellular signaling molecules, termed ligands, bind to specific receptor proteins on the cell's surface, initiating a physical and/or chemical reaction that is propagated within the cell by interacting proteins. Proteins involved in the signal transmission process form signaling ***pathways*** and these pathways in turn, assemble into complex networks that control cellular function (Pawson and Nash 2000). Misregulation in signaling networks has been observed in diseases such as cancer, muscular dystrophy, Alzheimer's, and Huntington's disease (Gonzalez M.W.;, Kann, M.G. 2012). Increased knowledge of protein interactions in signaling networks will lead to greater insight into the nature of those diseases and to therapeutic advances.

Protein interactions may be direct (physical), where two proteins directly bind to each other, or indirect (functional), where no direct contact takes place between the interacting partners. This work focuses on direct physical PPIs.

Figure 1.1.1: Cellular signal transduction involving STATs that affects gene regulation.



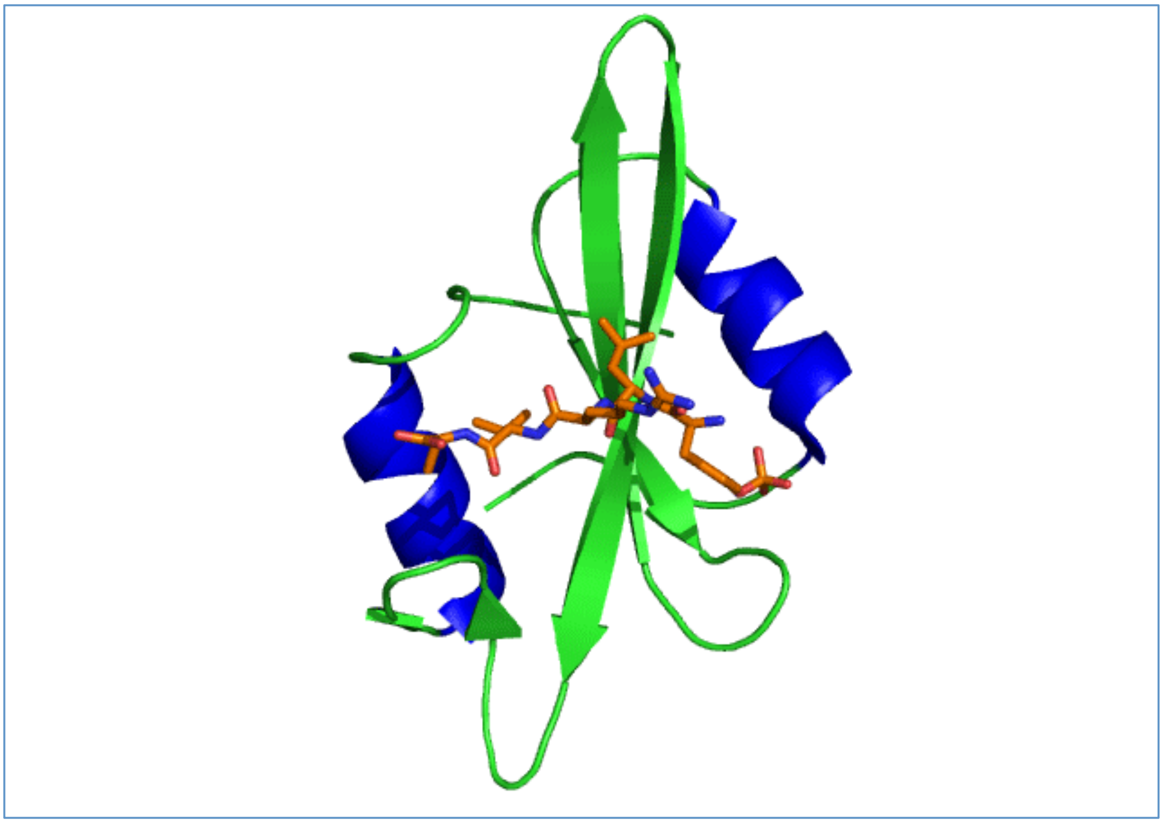
An external stimulator binds to TGF-beta receptors on the surface on the cell, initiating a series of interaction events that are propagated within the cell to invoke a response (Darnell, 1997).

http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&id=219&Itemid=67

1.2 – Peptide Recognition Modules

Many protein interactions depend on the activity of peptide recognition modules (PRMs). PRMs are globular protein domains that mediate interactions by binding to specific, short, linear regions of other proteins (Sidhu et al. 2014). Well-known examples of PRMs are SH2 and SH3 (Src homology 2 and 3) (Musacchio et al. 1992, Mayer et al. 2001, Pawson et al 2001), PDZ (**P**ostsynaptic Density 95 (PSD-95); ***d****iscs large* (DLG) and *zonula occludens*-1 (ZO-1)) (Kennedy 1995, Doyle et al. 1996, Kim et al. 2004, Tonikian et al. 2008), PTB (phosphotyrosine-binding) (Zhou et al. 1995), and WW (named after the presence of two conserved tryptophan residues, abbreviated as 'W') (Bork and Sudol 1994, Sudol M 1996) domain family members. Each of these domains folds into well-characterized structures typical of that domain and recognizes specific peptide motifs (Pawson 2006), **[Fig. 1.2.1-1.2.4]**. For instance, the WW domain is a 38 amino acid-long domain, which contains two conserved tryptophan residues and binds proline-rich motifs (Nguyen et al. 1998, Wintjens et al. 2001). SH2 and PTB domains bind to phosphorylated tyrosine residues, while SH3 and WW domains recognize proline-rich peptides (Zhou et al. 1995, Kim and Sheng 2004, Pawson 2006). Members of these PRM families are involved in a variety of regulatory cellular processes (Pawson 2003), and mutations in them lead to misregulated pathways important in diseases such as cancer, Alzheimer's and Huntington's disease (Pawson 2000).

Figure 1.2.1: SH2 Domain

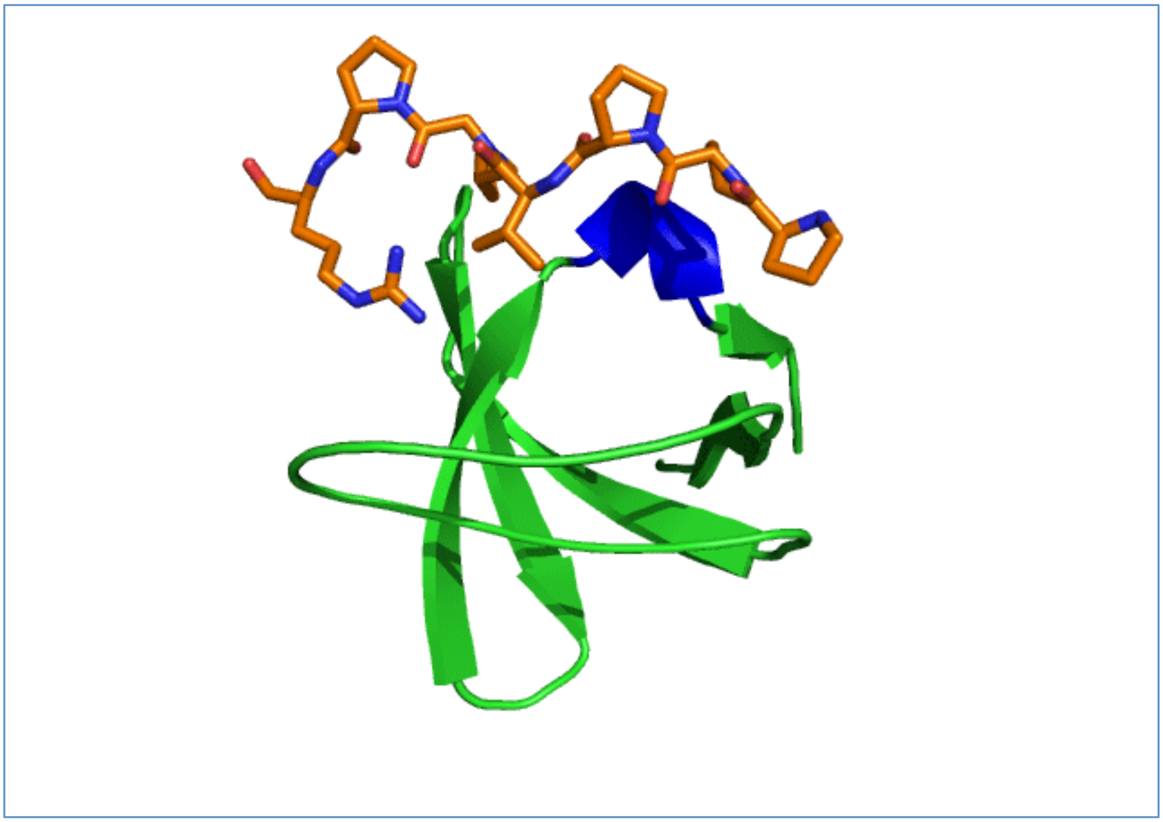


SH2 domain of v-src bound to a pYRLV peptide ligand

http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=178

SH2 domains contain a central anti-parallel beta-sheet (green) surrounded by two alpha-helices (blue). They bind phosphotyrosine (pY) peptides and are found in a variety of adaptor, scaffold and kinase proteins.

Figure 1.2.2: SH3 Domain

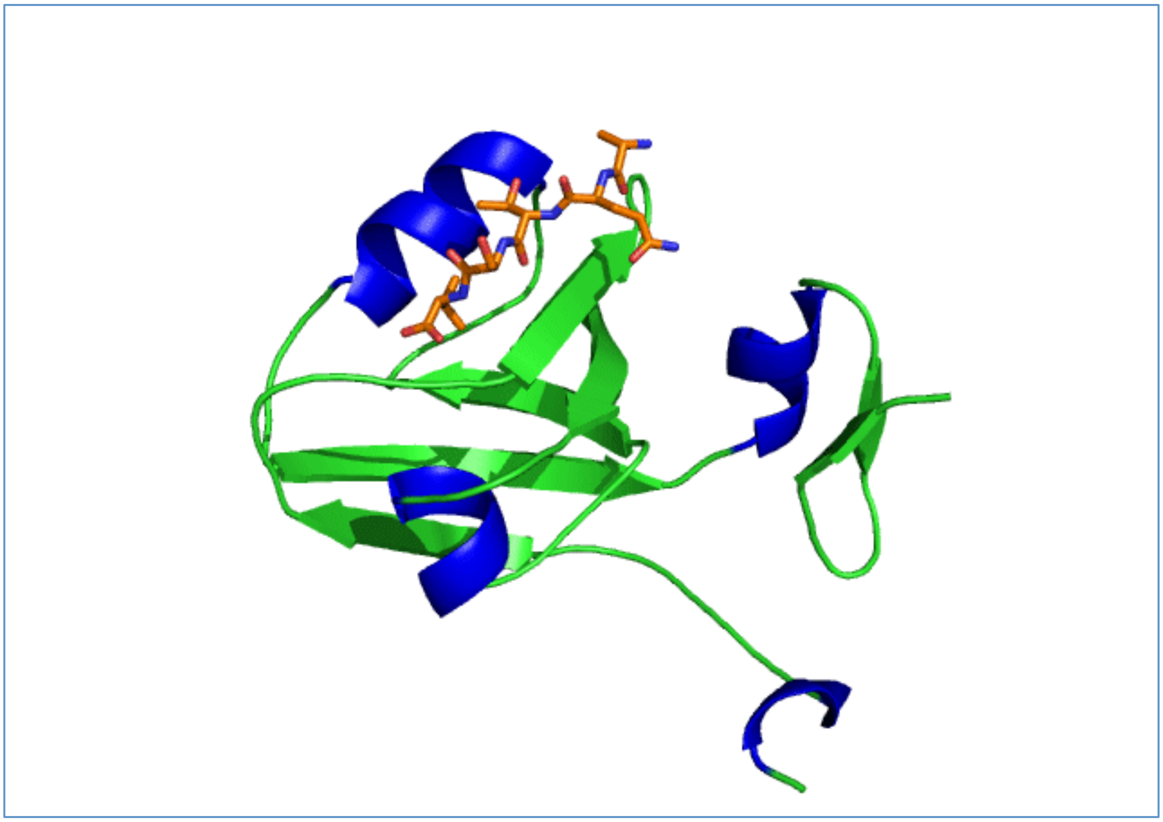


A Sem5 C-terminal SH3 (Src homology 3) domain complexed to the mSos-derived sequence PPPVGPRRR (Pawson 2005)**.**

http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=179

The SH3 domain contains five anti-parallel beta strands (green). The binding site (orange) forms a hydrophobic patch that contains a cluster of conserved aromatic residues and is surrounded by two charged and variable loops.

Figure 1.2.3: PDZ Domain

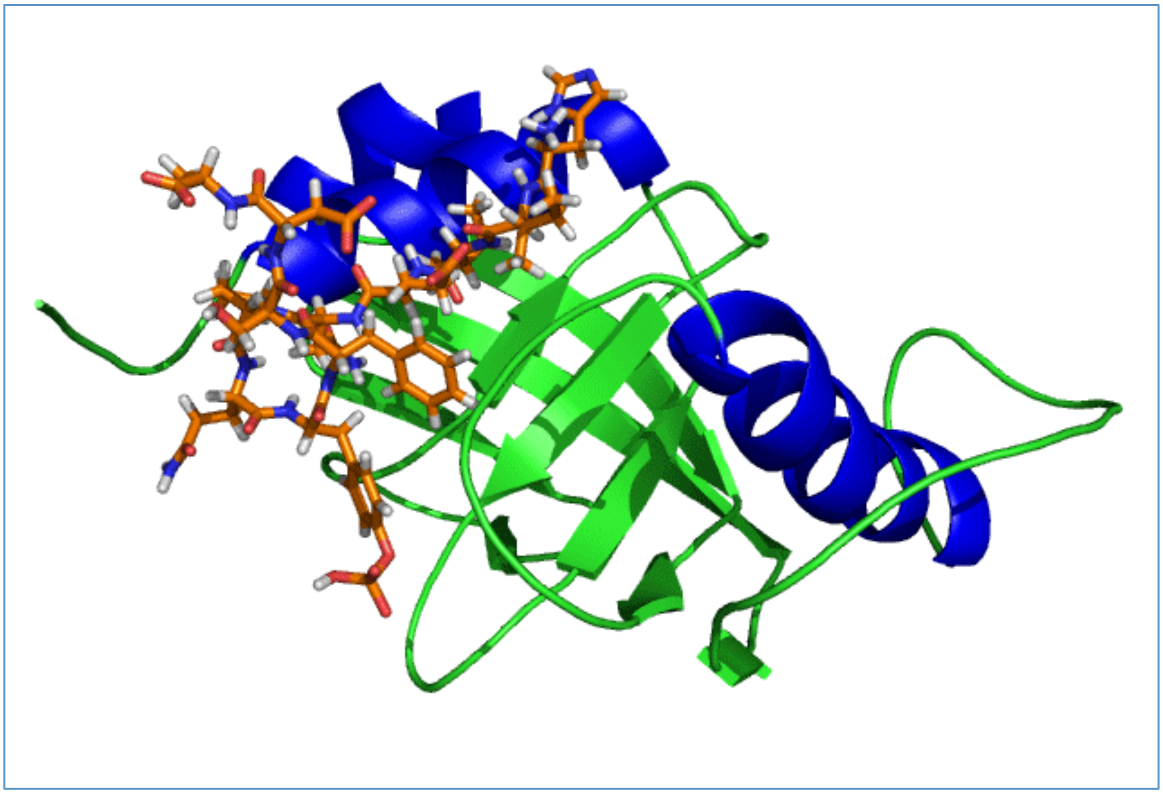


The third PDZ domain of PSD-95, bound to a TKNYKQTSV peptide (Pawson 2005)**.**

http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=168

PDZ domains are composed of approximately 80-90 amino acid (AA) residues and contain two alpha-helices (blue) and 5-6 beta strands (green). The binding site is a hydrophobic cleft that binds the peptide's carboxylate group.

Figure 1.2.4: PTB Domain



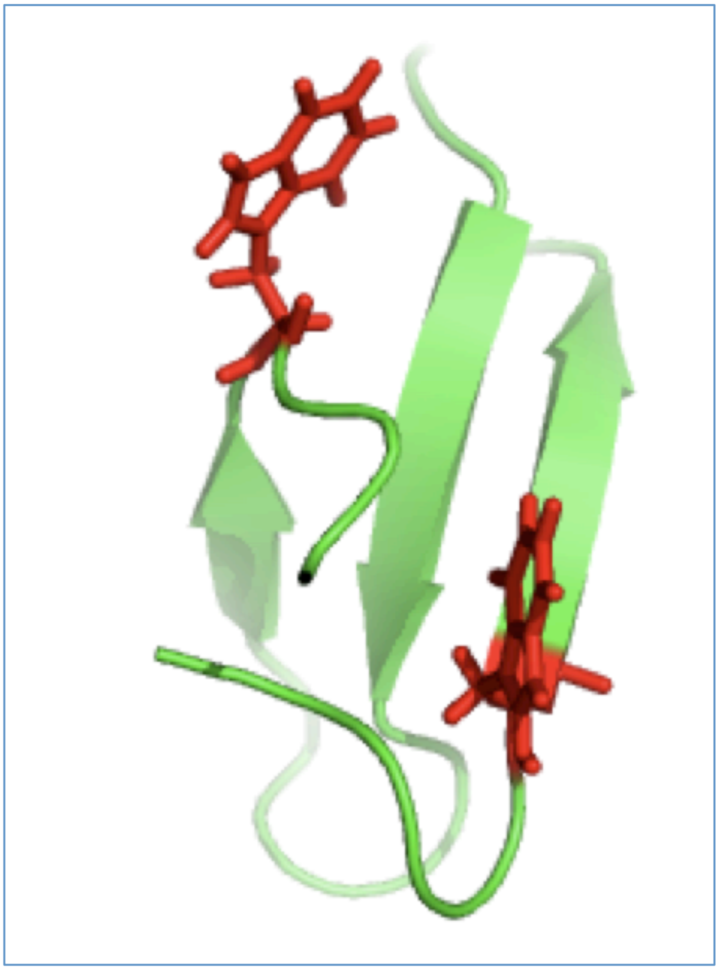
PTB domain of Shc complexed to a HIIENPQpYFSDA peptide –
http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=170

PTB (phospho-tyrosine binding) domain is composed of approximately 100-150 AA residues and binds phosphorylated proline-rich motifs (NPXpY). It contains two alpha-helices (blue) and 6 beta sheets (green).

1.3 – WW Domains

The WW domain (also known as WWP), described by Bork and Sudol in 1996, is a protein domain with two highly conserved tryptophan (‘W’) residues (Bork and Sudol 1994, Sudol 1996), [**Fig. 1.3.1**]. The WW domain recognizes and binds to proline-rich motifs (Sudol 1996, Pawson and Nash 2003, Ingham et al. 2005). It is composed of approximately 38 amino acid residues and folded into a three-stranded beta-sheet structure (Sudol 1996). Proteins containing this domain are involved in signal transduction in pathways such as the Hippo tumour suppressor pathway (Salah and Aqeilan 2011, Yu and Guan 2013), and mutations in the domain lead to misregulations that have been implicated in diseases such as cancer in mammals (Kodaka and Hata 2015). For this reason, the WW domain has been of overlapping interest in the Pawson and Bader laboratories; however, extensive research has not been done on them as much as on SH2 and SH3 domains. For these reasons, my thesis focuses on the WW domain.

Figure 1.3.1: WW Domain



Pin1 WW domain -
http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=191

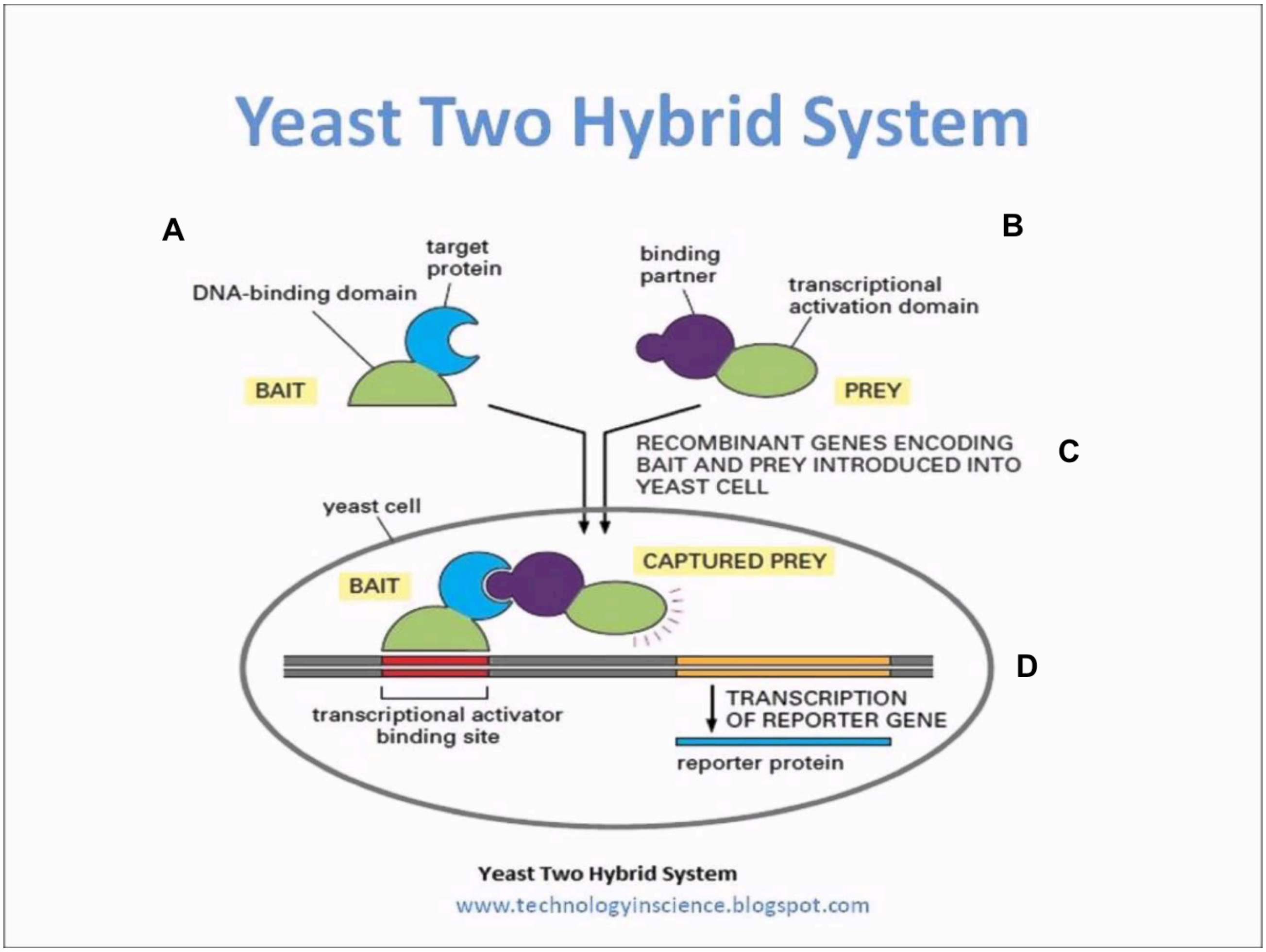
The WW domain is a 38-AA unit that folds into a 3-stranded beta sheet structure and binds proline-rich motifs. The name 'WW' illustrates that the domain contains two conserved tryptophan (Trp or 'W') residues within its consensus sequence.

1.4 – Experimental PPI identification

Cellular protein interactions can be identified using several experimental techniques. These techniques include tandem affinity purification (TAP) (Puig et al. 2001), yeast two-hybrid (Y2H) (Fields and Song 1989, Ito et al. 2001, Walhout and Vidal 2001, Brückner et al. 2009), co-immunoprecipitation (Co-IP) (Hall 2004, Isono and Schwechheimer 2010), peptide arrays (Wu et al. 2007, Katz et al. 2011, Amartely et al. 2014) or phage display (Kay and Castagnoli 2003, Kokoszka and Kay 2015).

The yeast-two hybrid (Y2H) method (Fields and Song 1989) detects PPIs based on the assembly of a transcription factor (TF) and the subsequent activation of selected "reporter" genes (Auerbach D, Stagljar I 2005). A "bait" and a "prey" protein hybrids are prepared in yeast. The "bait" is fused to the reporter gene's TF's DNA-binding domain (DBD), while the "prey" is fused to the TF's activation domain (AD). If the bait and prey interact when expressed in a yeast cell that contains a specific "reporter gene", the interaction brings the AD and DBD into close proximity, resulting in a functional TF, which, in turn triggers the expression of the reporter gene [**Fig. 1.4.1**]. Hence, the reporter gene is used as an indicator of interactions between the 'bait' and 'prey' proteins.

Figure 1.4.1: Yeast Two-Hybrid (Y2H) PPI detection method

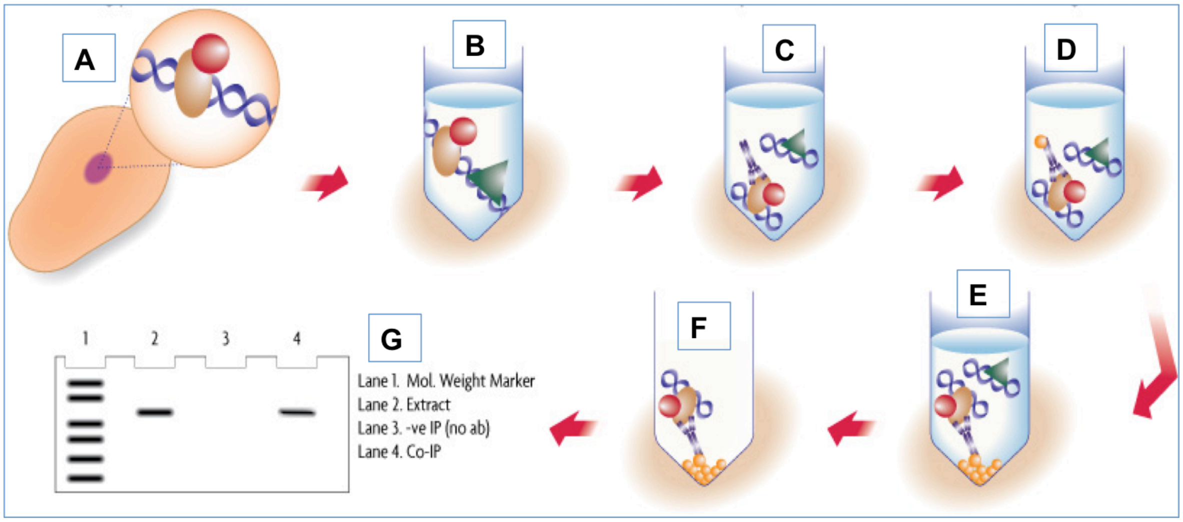

Yeast Two Hybrid System
A
B
target protein
DNA-binding domain
BAIT
binding partner
transcriptional activation domain
PREY
RECOMBINANT GENES ENCODING BAIT AND PREY INTRODUCED INTO YEAST CELL
C
yeast cell
CAPTURED PREY
BAIT
D
transcriptional activator binding site
TRANSCRIPTION OF REPORTER GENE
reporter protein
Yeast Two Hybrid System
www.technologyinscience.blogspot.com

The Yeast Two-Hybrid (Y2H) PPI detection method. A) A "bait" is prepared, by fusing the target protein (P1) to the DBD of the TF. B) A "prey" is prepared, by fusing a potential binding partner (P2) to the AD of the TF. C) Bait and prey are placed inside a yeast cell that contains a reporter gene. D) Proteins P1 and P2 interact, creating a functional TF, which activates transcription of the reporter gene.

(Photo from www.technologyinscience.blogspot.com)

Co-immunoprecipitation (Co-IP) is an extension of the immunoprecipitation (IP) protocol, commonly used for protein detection. IP targets the protein in question (antigen) with an antibody that the protein has a known affinity for and pulls down (precipitates) the antigen-antibody complex using an immobilizing agent, such as an Ab-binding protein, on a beaded surface. Proteins not precipitated on the beads are washed away, and the protein in question is detected using gel electrophoresis followed by Western blot. The Co-IP technique pulls down the bait protein along with its interacting partners and, for example, uses mass spectrometry to identify the bait protein. (Thermo Fisher Scientific Inc. 2015) [**Fig. 1.4.2**]. In the Tandem Affinity Purification (TAP) technique, the TAP tag is used instead of direct antibodies to label and later detect the bait proteins. Usually, the bait and prey are then identified by mass spectrometry. (Puig et al. 2001)

Figure 1.4.2: Co-immunoprecipitation (Co-IP) PPI detection method.



Co-IP schema. A) Interacting proteins *in vivo*. B) Proteins in question are extracted from the cell and placed *in vitro* in an extract of low-salt buffer with enzymatic shearing to protect the protein complexes C) Antibody with known affinity to one of the interacting proteins (bait) is added *in vitro* D) Antibody binding beads are added E) Solution is washed, and proteins in question are immunoprecipitated on the beads F) Proteins of interest are collected G) A Western blot is performed to analyze the immunoprecipitated PPI using an Antibody against the bait's interactor

Picture taken from http://www.activemotif.com/images/products/coip_flowchart_big.jpg

1.5 – Computational PPI Predictions

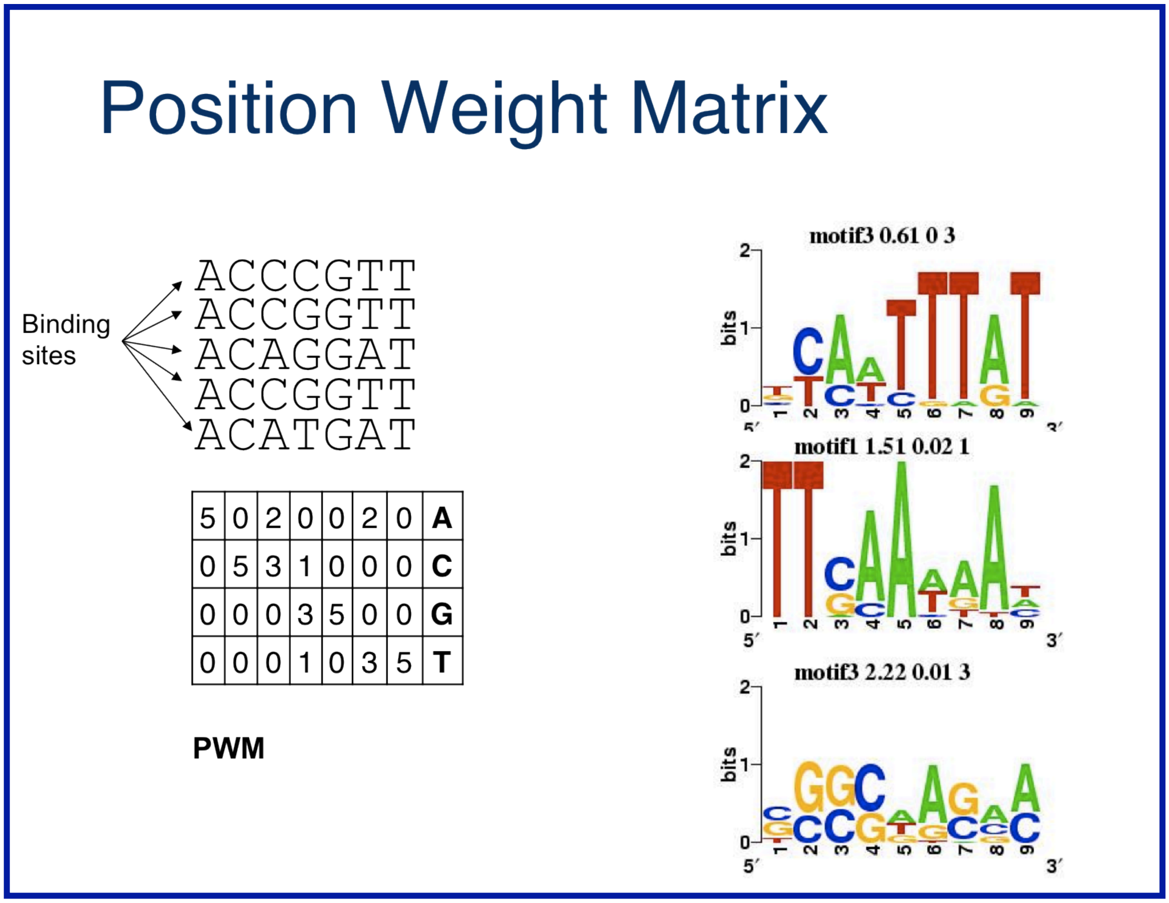
1.5.1 – The need for computational predictions

Despite their accuracy, experimental methods for PPI identification remain costly and time-consuming. Y2H and affinity-based techniques have shown 40%-80% false negative rate and around 12% false positive rate, resulting in only partially complete interactome maps (Venkatesan et al. 2009). Moreover, large-scale results for thousands of samples often do not answer specific questions related to a particular protein (Leser and Hakenberg 2005). The Y2H method, for instance, cannot reveal interactions between more than two proteins (Berggård et al. 2007). Complementing wet-lab procedures with computational predictions can increase the financial and temporal effectiveness of wet-lab methodologies in identifying high quality protein interaction networks.

1.5.2 – PPI prediction methods

Computational predictions are made by applying bioinformatics and machine-learning algorithms to training sets obtained from large-scale experiments, such as peptide arrays or phage display. Alternatively, prediction algorithms may extract information on protein-protein interactions from high-volume collections of published data that do not directly identify protein interactions, but nonetheless are correlated with protein interactions. A simple method for predicting PRM-dependent protein interactions involves position-weight matrices or position-specific scoring matrices (PWMs or PSSMs) (Sinha 2006, Kerpedjiev et al. 2014). In this case, PWMs capture the probability of an amino acid residue to occur at a specific position in a peptide that is predicted to bind to a PRM, which can be visualized as a sequence logo [**Fig. 1.5.1**]. Other computational PPI prediction methods include high-throughput sequence-based approaches (Chen and Jeong 2009, Liu et al. 2012, You et al. 2014), structure-based approaches (Hosur 2012), function-based approaches (Schlicker et al. 2006, Wang et al. 2007), chromosome proximity (Vijaykumar and Vishal 2013), gene clustering (Lee and Sonnhammer 2003), *in-silico* two-hybrid, phylogenetic tree, phylogenetic profile, and gene expression-based approaches (Rao et al. 2014).

Figure 1.5.1: Position-weight matrix and sequence logo



Position Weight Matrix

Binding sites

ACCCGTT
ACCGGTT
ACAGGAT
ACCGGTT
ACATGAT

5	0	2	0	0	2	0	**A**
0	5	3	1	0	0	0	**C**
0	0	0	3	5	0	0	**G**
0	0	0	1	0	3	5	**T**

PWM

Saurabh Sinha: ***"Counting position weight matrices in a sequence & an application to discriminative motif finding"*** Computer Science, University of Illinois, Urbana-Champaign

1.7 – The downside of computational predictions

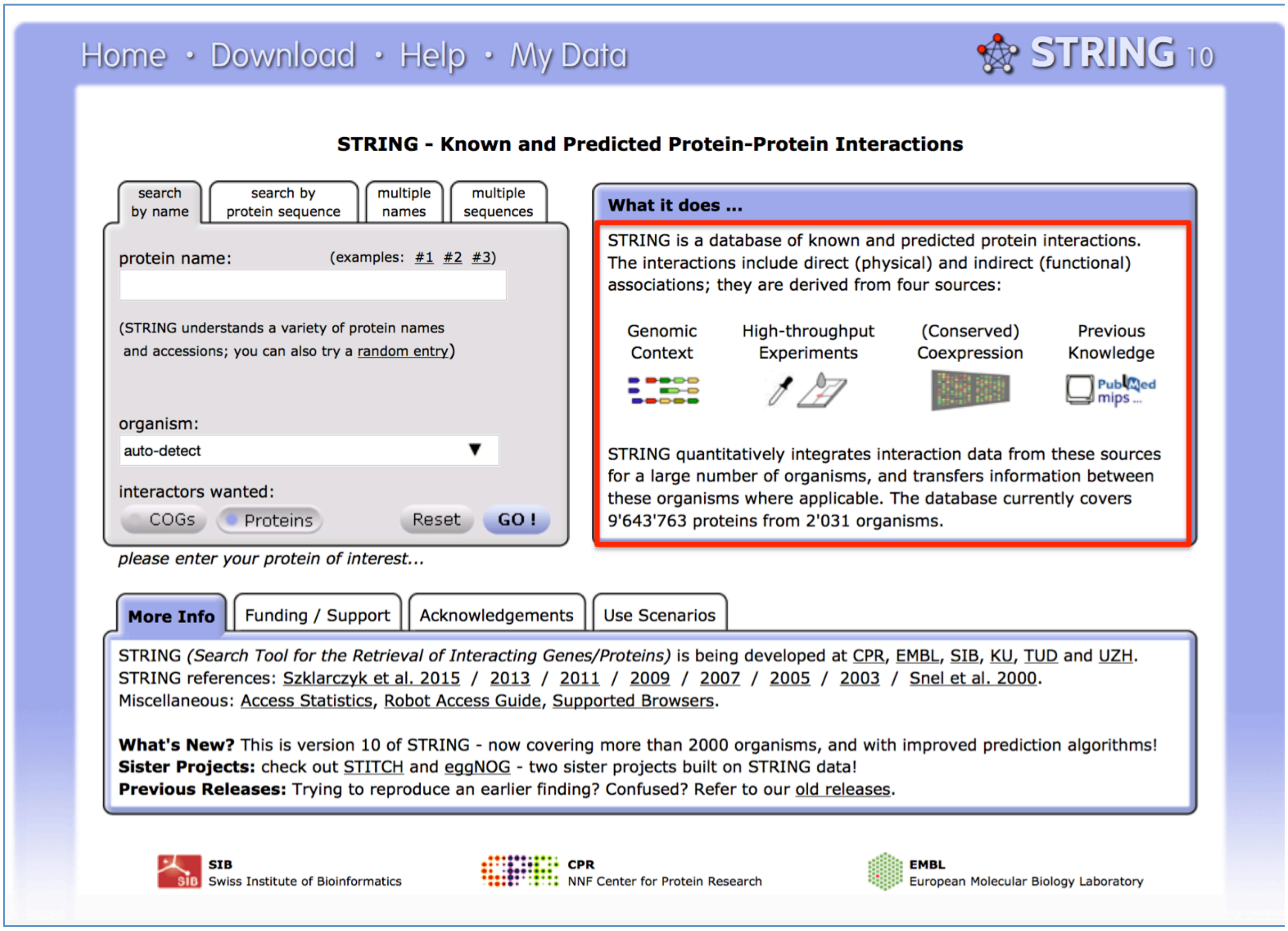
Predictions made *in silico* contain many false positives, as bioinformatics methods such as PWMs do not take into account the biological context that would make these interactions possible *in vivo*. Even though computational methods can identify the ***potential*** of proteins to interact, in reality conditions must be met within the cell in order to enable interactions. These conditions are termed ***biological context***. Examples of biological context include co-localization (presence in the same location in the cell within reasonable proximity of each other), co-expression (presence in the cell at the same time), favorable conformation that permits interaction, or accessibility of interacting regions on the proteins' surfaces. Hence, a solid computational prediction model considers 'real' interactions, i.e. possible under certain biological conditions ("cellular context"), to reduce the number of false positives and discard predicted interactions that are not likely to occur *in vivo*.

1.6 – Sources of evidence for PPIs

Prediction datasets in a specific biological context are termed "sources of evidence" (or "evidence sources") for protein interactions. Co-localization, co-expression, surface accessibility, pathway co-occurrence, correlated mutations, and evolutionary conservation are all examples of evidence sources. Integrating data from multiple independent supporting evidence sources improves prediction accuracy, as seen in tools such as PrePPI (Zhang et al. 2012) that combines three-dimensional structural and functional information. Repositories such as STRING (Snel et al. 2008, Szklarczyk et al.

2015) [**Fig. 1.6.1**], PIPs [**Fig. 1.6.**2] and other contain information on PPIs from different evidence sources.

<u>Figure 1.6.1</u>: STRING database



STRING - Known and Predicted Protein-Protein Interactions

protein name: (examples: #1 #2 #3)

(STRING understands a variety of protein names and accessions; you can also try a random entry)

organism: auto-detect ▼

interactors wanted:

please enter your protein of interest...

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

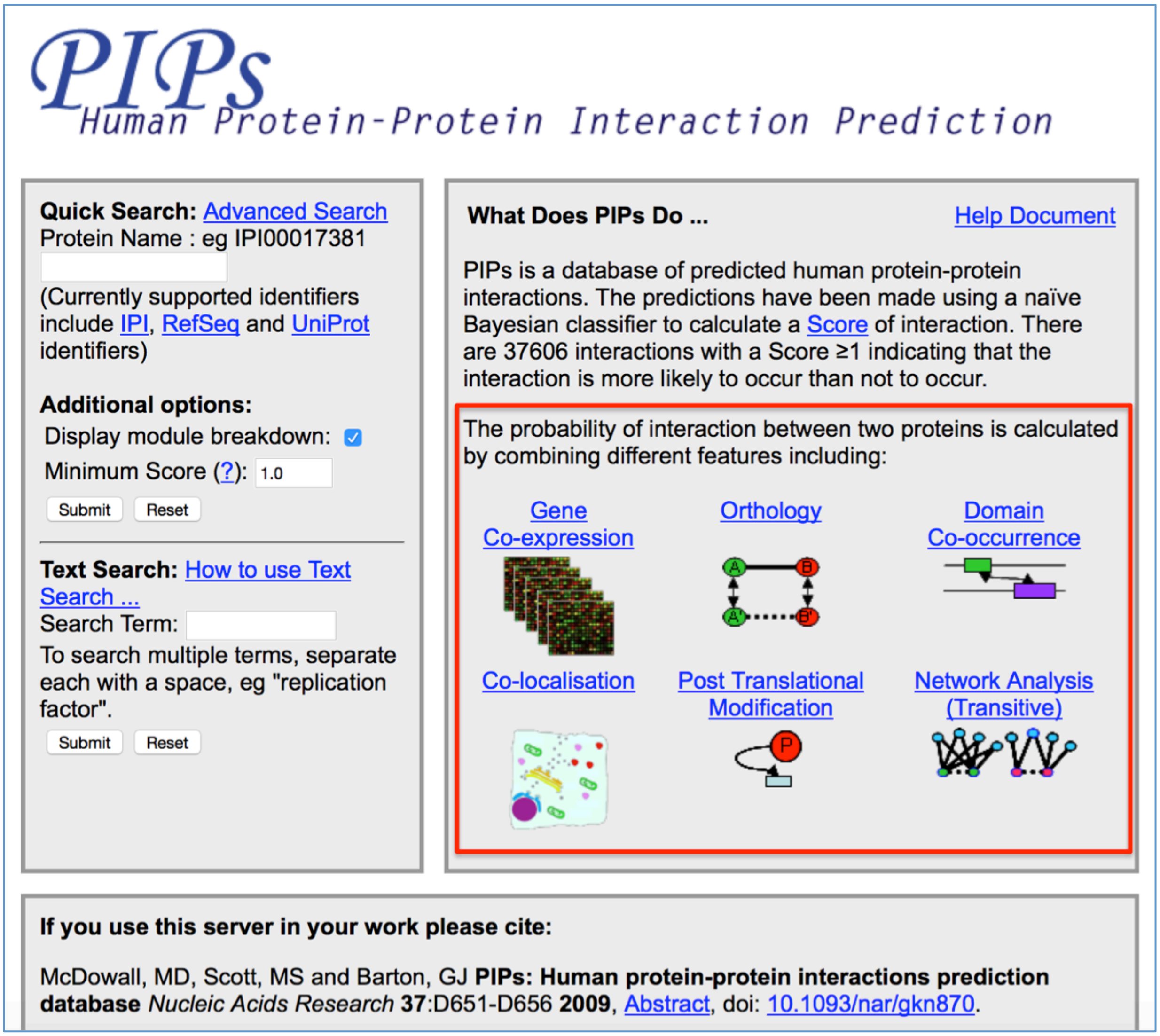
STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 9'643'763 proteins from 2'031 organisms.

STRING *(Search Tool for the Retrieval of Interacting Genes/Proteins)* is being developed at CPR, EMBL, SIB, KU, TUD and UZH.
STRING references: Szklarczyk et al. 2015 / 2013 / 2011 / 2009 / 2007 / 2005 / 2003 / Snel et al. 2000.
Miscellaneous: Access Statistics, Robot Access Guide, Supported Browsers.

What's New? This is version 10 of STRING - now covering more than 2000 organisms, and with improved prediction algorithms!
Sister Projects: check out STITCH and eggNOG - two sister projects built on STRING data!
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our old releases.

STRING (Snel et al. 2008, Szklarczyk et al. 2015) **is a repository of both known and predicted physical and functional PPIs. Predicted interactions are obtained from different evidence sources.**

Figure 1.6.2: **The PIPs repository**



PIPs
Human Protein-Protein Interaction Prediction

Quick Search: Advanced Search
Protein Name : eg IPI00017381

(Currently supported identifiers include IPI, RefSeq and UniProt identifiers)

Additional options:
Display module breakdown: ☑
Minimum Score (?): 1.0

Text Search: How to use Text Search ...
Search Term:
To search multiple terms, separate each with a space, eg "replication factor".

What Does PIPs Do ... Help Document

PIPs is a database of predicted human protein-protein interactions. The predictions have been made using a naïve Bayesian classifier to calculate a Score of interaction. There are 37606 interactions with a Score ≥1 indicating that the interaction is more likely to occur than not to occur.

The probability of interaction between two proteins is calculated by combining different features including:

If you use this server in your work please cite:

McDowall, MD, Scott, MS and Barton, GJ **PIPs: Human protein-protein interactions prediction database** *Nucleic Acids Research* **37**:D651-D656 **2009**, Abstract, doi: 10.1093/nar/gkn870.

The PIPs database is a PPI prediction repository that combines predicted data from different evidence sources.

1.7 – Text Mining

Text mining, or information extraction, is automated extraction of structured information from text using bioinformatics and machine-learning algorithms. It can help to quickly uncover hidden or previously unknown information in high volumes of unstructured text without human intervention. The most prominent example of this type of text mining is a web search engine such as Google, which extracts information based on keywords from a repository of websites (Hill and Lewicki 2007). In biology, text mining of large volumes of published data is applied to assist biologists in quickly uncovering information. An example of the application of text mining in biology is the 'Related Articles' function in PubMed, where a content similarity algorithm is used to retrieve articles similar to the search term (Lin and Wilbur 2007).

At the basis of text mining is Named Entity Recognition (NER) – a strategy for identifying the terms (entities) in question. NER is a technique to recognize concepts in text that follow a selected form. Statistical methods and machine-learning techniques, such as support vector machines (SVMs) (Takeuchi and Collier 2002), Hidden Markov Models (HMM), Maximum Entropy (ME), and Conditional Random Fields (CRF are then applied to extract and analyze the NER results.

This Masters work describes the use of text mining of PubMed abstracts as a novel evidence source for WW domain-mediated PPIs.

1.8 – Precision and Recall

To evaluate the performance and utility of computational prediction algorithms, it is necessary to determine how well the algorithm retrieves results that are true and how well it discards results that are false. Two standard statistical measurements to compute these ratios are precision and recall. Precision refers to the fraction of true positives out of all predictions. Recall describes the fraction of true positives out of everything in the benchmark. 'True positives' refers to prediction results that have been proven true – in our case, a true positive would be a predicted interaction that has been experimentally validated and shown to occur *in vivo*. Correspondingly, 'false positives' refer to results that have been predicted true by the algorithm, but are actually false (in our case – predicted interactions that have not been experimentally verified). A repository of known, experimentally validated interactions serves as a 'benchmark'. In this work, the benchmark is the iRefIndex version 9 repository of PPIs obtained using different experimental methods.

1.9 – Thesis Outline

The first chapter of the thesis describes my work on gene name co-occurrence in PubMed abstracts. I programmatically searched these abstracts for co-occurrence of gene symbols of WW domain-containing proteins and evaluated these predictions statistically, by setting a cutoff for the number of abstracts and plotting a precision-recall graph. I then proceeded to examine the Medical Subject Heading (MeSH) terms indexed in those articles and compared the results to previous findings. This work is outlined in Chapter

2. Finally, I refined the co-occurrence prediction approach by introducing a controlled vocabulary into the search (Chapter 3). All results are presented graphically in figures throughout the thesis. Python scripts used in the process are presented in the Appendix.

CHAPTER 2. GENE NAME CO-OCCURRENCE

Co-occurrence based methods are one common technique in text mining to predict and construct PPI networks (Jenssen et al. 2001, Cohen and Hunter 2008). Concepts that are mentioned within the same unit of text, such as sentence or paragraph, are predicted to also have a biological affiliation (Krallinger et al. 2008). In protein interaction predictions, the co-occurring terms may be gene names, gene symbols, GO terms (Jain and Bader 2010) or MeSH terms (Jenssen et al. 2001). Co-occurrence can be used as computational evidence of biological association. In the STRING repository, co-occurrence of genes serves as an indication of functional relation (Snel B 2008).

Research has been done in the Bader lab to predict PRM-mediated PPIs involving SH3 and PDZ domain family members, using GO terms as an evidence source. Jain et al. (2010) predicted that proteins with similar GO gene function annotation also potentially interact in vivo. The procedure described below identifies co-occurring human gene symbols in PubMed abstracts to be used as an evidence source for WW-mediated PPI predictions. The confidence measure is the number of co-occurrences for protein names in question; i.e. the higher the number of abstracts where these protein names co-occur, the more likely these proteins are to interact *in vivo*.

2.1 – MATERIALS AND METHODS

The Python programming language was selected as the language for writing custom text mining scripts, since it is a language that I am extensively familiar with, as well as one of the preferred languages for bioinformatics analyses, with its built-in `Bio` and `EUtils` libraries. In addition, Python is open-source, lightweight for installation and configuration on UNIX-like platforms, easy to learn and execute.

Scripts written in Python are executed in UNIX-like environments, such as Linux or Mac.

As an auxiliary resource, a custom database [**Appendix A**] was constructed, using the MySQL database management system (DBMS), by virtue of its being open-source, freely available, flexible for installation on different operating systems (OS), including Linux and Mac, and my extensive familiarity with it.

The **iRefIndex** (Turner et al. 2010) database was used as the benchmark, since it contains information on both predicted and experimentally verified PPIs, which can be either direct or indirect, physical or chemical, and detected using various methods. iRefIndex integrates information on PPIs from different databases, including BIND, BioGrid, DIP, HPRD, MPPI, OPHID and more. The iRefWeb web interface lets the user select the number of databases where PPIs were observed (1 or more, 2 or more, 3 or more). iRefIndex was selected as the benchmark set, as it includes interactions from different databases obtained using different experimental methods. Other PPI repositories, such as HPRD, do not include data from multiple evidence sources, and, hence, provide less information than iRefIndex.

2.2 – PROCEDURE

2.2.1. Datasets:

The text mining process started with downloading the following datasets:

- A list of all approved official HGNC symbols for every protein-coding gene in the human genome (**32717** in total) in text format – **Set A**

- A list of **50** gene symbols of WW domain-containing proteins, downloaded from Ensembl (http://ensembl.org) using the Biomart query system (Smedley et al. 2009, Zhang et al. 2011) in text format – **Set B** **(Appendix B)**

- A set of all PubMed abstracts for every WW-containing protein in **Set B**, excluding DMD and ITCH. The abstracts can be retrieved using the built-in '*esearch*' and '*efetch*' methods of the Entrez Programming Utilities Python package (**EUtils**) (National Center for Biotechnology Information (US) 2008), or, alternatively, downloaded directly from PubMed. DMD and ITCH were excluded from the search, since the number of abstracts matching these terms exceeds 1000, and the script times out. (**Appendix C**). **Appendix D** contains a Python script to parse the XML results and store them in the database.

- A list of **1722** interacting protein pairs, in text format, downloaded from the iRefWeb (iRefIndex database version 9) in full MITAB format.

  To obtain the benchmark dataset, the iRefWeb site http://wodaklab.org/iRefWeb/search/index was searched using the following parameters:

  - **Source database:** ANY
  - **Organism:** Single organism interaction, *Homo sapiens*
  - **Nature of Interactions:** Pairwise, experimental, physical
  - **Number of Publications:** 1 or more publications

The interacting protein pairs – were downloaded in MITAB format. MITAB is the standard format for biological data exchange, as specified by the Human Proteome Organization (HUPO) Proteomics Standard Initiative (PSI).

The downloaded MITAB file contains the following information:
(http://psidev.sourceforge.net/molecular_interactions/xml/doc/user/)

Columns 1 and 2:

```
uidA  uidB
```

Unique identifiers, mainly ‘UniProtKB’, of the interacting proteins:

e.g. `uniprotkb:Q05193      uniprotkb:P60880`

Columns 3 and 4:

```
altA  altB
```

List of alternative identifiers of the interactors, separated by |

e.g.

```
uniprotkb:P60880edgetypeuniprotkb:Q05193|refseq:NP_004399|entrezgene/locuslink:1759|rogid:uiP8CXhKWQaP2GIAZULJTLqwGLs9606|irogid:4370876
```

Columns 5 and 6:

```
aliasA     aliasB
```

List of aliases for the interacting proteins, separated by |.

e.g.

```
uniprotkb:DYN1_HUMAN|entrezgene/locuslink:DNM1|crogid:uiP8CXhKWQaP2GIAZULJTLqwGLs9606|icrogid:4370876
```

This is the input I used for my benchmark, since it contains values in the form GENESYMBOL_HUMAN. In this work, all values that contain the term ‘HUMAN’ were extracted from the MITAB file; built-in UNIX commands and VI editor were used to remove all other information.

Column 7:

`method` – Interaction detection methods, separated by |

Column 8:

`author` – Author(s) of publications where this interaction was shown, separated by |

Column 9:

`pmids` – PubMed IDs of publications where this interaction was shown, separated by |

Columns 10 and 11:

`taxa taxb` – NCBI taxonomy identifiers for interactors A and B

The remaining columns are internal iRefWeb identifiers and scores, not used in this work.

Alternatively, interaction data may be downloaded from iRefWeb in MITAB-lite format, which contains condensed information. Since it contains no protein names, only internal identifiers, it was not used as a benchmark in this work.

The benchmark set was downloaded in the full MITAB format, and all information not pertaining to WW domain-containing protein interaction was removed using the Python programming language (**ww_benchmark.py**), as well as standard UNIX commands and the built-in **VI** editor.

The final benchmark set contains **1722** interacting protein pairs (this number has likely increased since 2012) **ww_benchmark.tsv** – **Appendix E**.

2.2.2. Text mining:

Once the datasets were downloaded, Python scripts were written and executed to extract protein names from PubMed abstracts. Proteins whose names were mentioned in the same abstract were predicted to interact. The interacting pair, along with the PubMed publication IDs (PMIDs) of the abstracts where the interaction was found, were recorded in a tab-delimited file.

Scripts:

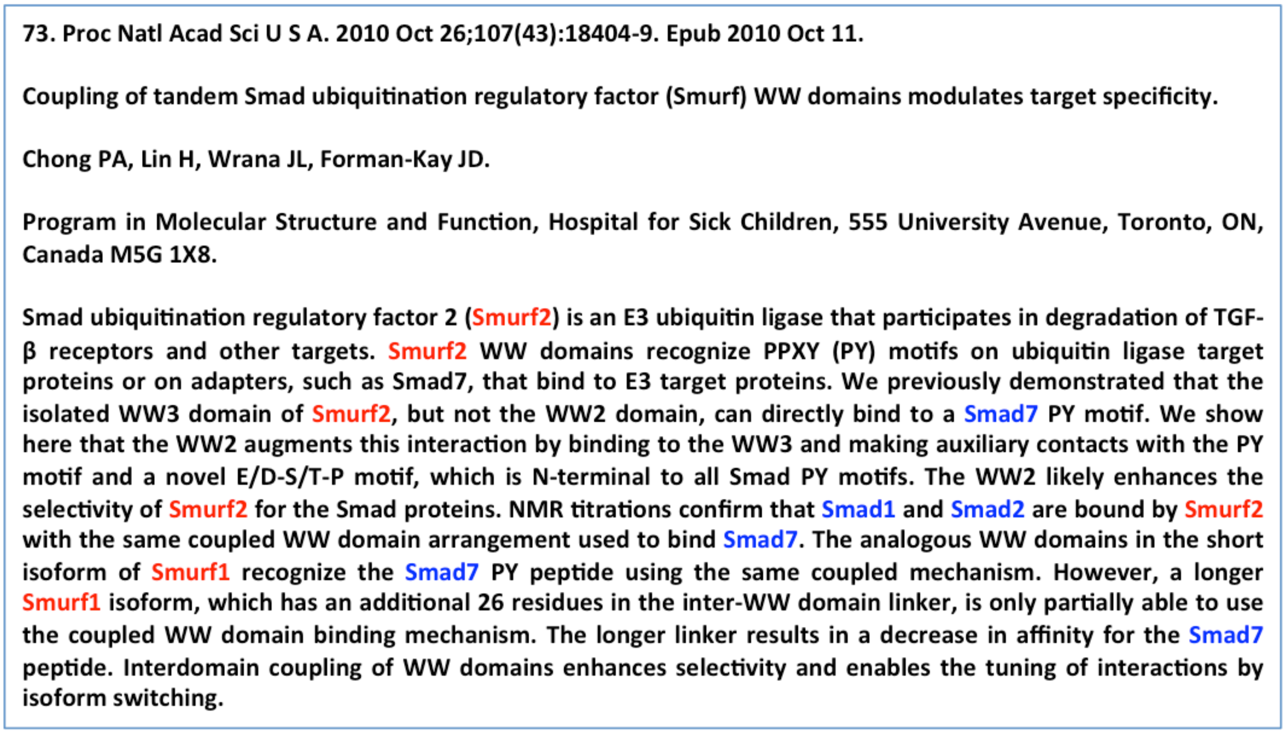
- **co_occurrence.py**: **(Appendix F)**

  Contains the script to identify PubMed abstracts where a WW-containing protein and any other proteins from the set of HUGO gene symbols co-occur (**approved_hgnc_symbol_biomart.txt**). The pair of protein names is recorded as a predicted interaction, along with the PMID of the abstract where they were found.

Recorded interactions were grouped by the number of abstracts where the interaction was found and verified against the benchmark (**Appendix G**) to reveal TPs and FPs. A predicted interaction that was also found in the benchmark was recorded as a TP. An interaction that was predicted but did not appear in the benchmark was recorded as a FP.

2.3 – RESULTS

Figure 2.3.1: Example of protein name co-occurrence in abstracts.



73. Proc Natl Acad Sci U S A. 2010 Oct 26;107(43):18404-9. Epub 2010 Oct 11.

Coupling of tandem Smad ubiquitination regulatory factor (Smurf) WW domains modulates target specificity.

Chong PA, Lin H, Wrana JL, Forman-Kay JD.

Program in Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G 1X8.

Smad ubiquitination regulatory factor 2 (Smurf2) is an E3 ubiquitin ligase that participates in degradation of TGF-β receptors and other targets. Smurf2 WW domains recognize PPXY (PY) motifs on ubiquitin ligase target proteins or on adapters, such as Smad7, that bind to E3 target proteins. We previously demonstrated that the isolated WW3 domain of Smurf2, but not the WW2 domain, can directly bind to a Smad7 PY motif. We show here that the WW2 augments this interaction by binding to the WW3 and making auxiliary contacts with the PY motif and a novel E/D-S/T-P motif, which is N-terminal to all Smad PY motifs. The WW2 likely enhances the selectivity of Smurf2 for the Smad proteins. NMR titrations confirm that Smad1 and Smad2 are bound by Smurf2 with the same coupled WW domain arrangement used to bind Smad7. The analogous WW domains in the short isoform of Smurf1 recognize the Smad7 PY peptide using the same coupled mechanism. However, a longer Smurf1 isoform, which has an additional 26 residues in the inter-WW domain linker, is only partially able to use the coupled WW domain binding mechanism. The longer linker results in a decrease in affinity for the Smad7 peptide. Interdomain coupling of WW domains enhances selectivity and enables the tuning of interactions by isoform switching.

Smurf1, Smurf2, Smad1, Smad 2 and Smad7 gene names co-occur in this abstract. These proteins are also listed as interacting partners in GeneMANIA (Mostafavi et al. 2008, Warde-Farley et al. 2010)**, as illustrated in Figure 2.3.2.**

Figure 2.3.2 has been obtained using the GeneMANIA visualization tool, which illustrates protein interactions of different types (direct or indirect). The strength of interactions corresponds to the thickness of lines connecting the proteins in question.

Figure 2.3.2: A diagram illustrating the interaction between Smurf1, Smurf2, Smad1, Smad2 and Smad7 proteins from the GeneMANIA prediction server (Warde-Farley et al. 2010) (http://www.genemania.org)

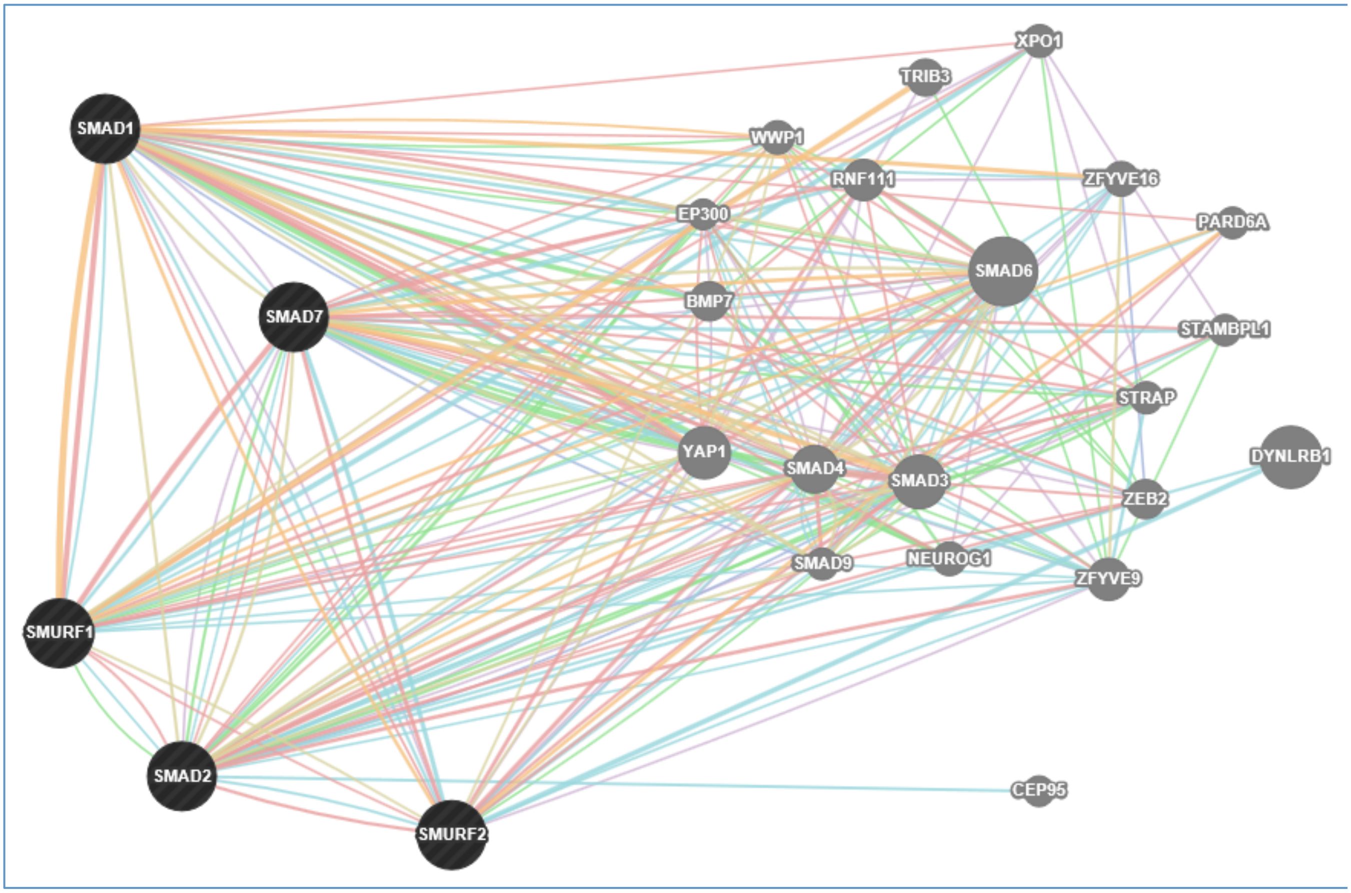


Figure 2.3.3: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Zhang et al. 2001).

> **Smad proteins are key intracellular signaling effectors for the transforming growth factor-β superfamily of peptide growth factors. Following receptor-induced activation, Smads move into the nucleus to activate transcription of a select set of target genes. The activity of Smad proteins must be tightly regulated to exert the biological effects of different ligands in a timely manner. Here, we report the identification of Smurf2, a new member of the Hect family of E3 ubiquitin ligases. Smurf2 selectively interacts with receptor-regulated Smads and preferentially targets Smad1 for ubiquitination and proteasome-mediated degradation. At higher expression levels, Smurf2 also decreases the protein levels of Smad2, but not Smad3. In *Xenopus*embryos, ectopic Smurf2 expression specifically inhibits Smad1 responses and thereby affects embryonic patterning by bone morphogenetic protein signals. These findings suggest that Smurf2 may regulate the competence of a cell to respond to transforming growth factor-β/bone morphogenetic protein signaling through a distinct degradation pathway that is similar to, yet independent of, Smurf1.**

Proc Natl Acad Sci U S A. 2001 Jan 30;98(3):974-9.

Regulation of Smad degradation and activity by Smurf2, an E3 ubiquitin ligase.
Zhang Y[1], Chang C, Gehling DJ, Hemmati-Brivanlou A, Derynck R.

PMID: 11158580

Figure 2.3.4: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Fukasawa et al. 2004).

Overexpression of transforming growth factor beta (TGF-**β**) has been shown to play pathogenic roles in progression of renal fibrosis, and the severity of tubulointerstitial fibrosis correlates better with renal function than the severity of glomerulosclerosis. Smad proteins are signaling transducers downstream from TGF-**β** receptors. Three families of Smad proteins have been identified: receptorregulated Smad2 and Smad3, common partner Smad4, and inhibitory Smad7 (part of a negative-feedback loop). We investigated Smad-mediated TGF-**β** signaling pathway and regulatory mechanisms of inhibitory Smad7 in unilateral ureteral obstruction (UUO) kidneys in mice, a model of progressive tubulointerstitial fibrosis. Compared with sham-operated kidneys, the level of Smad7 protein, but not mRNA, decreased progressively in UUO kidneys, whereas immunoreactivity for nuclear phosphorylated Smad2 and Smad3 and renal fibrosis were inversely increased. Furthermore, we demonstrated that both the degradation and ubiquitination activity of Smad7 protein were increased markedly in UUO kidneys compared with sham-operated ones. We also found that both Smurf1 and Smurf2 (Smad ubiquitination regulatory factors), which are E3 ubiquitin ligases for Smad7, were increased and that they interacted with Smad7 in UUO kidneys. Our results suggest that the reduction of Smad7 protein resulting from enhanced ubiquitin-dependent degradation plays a pathogenic role in progression of tubulointerstitial fibrosis.

Proc Natl Acad Sci U S A. 2004 Jun 8;101(23):8687-92. Epub 2004 Jun 1.

"Down-regulation of Smad7 expression by ubiquitin-dependent degradation contributes to renal fibrosis in obstructive nephropathy in mice."

Fukasawa H[1], Yamamoto T, Togawa A, Ohashi N, Fujigaki Y, Oda T, Uchida C, Kitagawa K, Hattori T, Suzuki S, Kitagawa M, Hishida A.

PMID: 15173588

Figure 2.3.5: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin et al. 2000)

Smads are important intracellular signaling effectors for transforming growth factor-beta (TGF-beta) and related factors. Proper TGF-beta signaling requires precise control of Smad functions. In this study, we have identified a novel HECT class ubiquitin E3 ligase, designated Smurf2, that negatively regulates Smad2 signaling. In both yeast two-hybrid and in vitro binding assays, we found that Smurf2 could interact with receptor-activated Smads (R-Smads), including Smad1, Smad2, and Smad3 but not Smad4. Ectopic expression of Smurf2 was sufficient to reduce the steady-state levels of Smad1 and Smad2 but not Smad3 or Smad4. Significantly, Smurf2 displayed preference to Smad2 as its target for degradation. Furthermore, Smurf2 exhibited higher binding affinity to activated Smad2 upon TGF-beta stimulation. The ability of Smurf2 to promote Smad2 destruction required the HECT catalytic activity of Smurf2 and depended on the proteasome-dependent pathway. Consistent with these results, Smurf2 potently reduced the transcriptional activity of Smad2. These data suggest that a ubiquitin/proteasome-dependent mechanism is important for proper regulation of TGF-beta signaling.

J Biol Chem. 2000 Nov 24;275(47):36818-22.

"Smurf2 is a ubiquitin E3 ligase mediating proteasome-dependent degradation of Smad2 in transforming growth factor-beta signaling."

Lin X[1], Liang M, Feng XH.

PMID: 11016919

Precision and Recall

A standard measure for evaluating the accuracy of bioinformatics methods is precision and recall computation. Precision, also known as ‘specificity’, is the fraction of true positives out of all predictions. Recall, also referred to as ‘sensitivity’, is the fraction of true positives out of everything in the benchmark. Precision and recall are computed according to the following formulas:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN),

‘TP’ and ‘FP’ represent the number of true positives and false positives in the prediction set.

In this work, predictions were grouped by the number of abstracts in which they were encountered. For every predicted interacting protein pair, the number of abstracts in which this prediction was encountered was also recorded. The higher the number of abstracts in which a given protein pair was encountered, the higher the likelihood that these proteins interact in vivo. Ideally, precision and recall would grow as the abstract cutoff increases.

Since the obtained precision and recall values are less than 1.0, they have been computed to 4 (four) significant figures, shown in the tables below.

Table 2.3.6: WW domain-containing protein interactions, grouped by the number of abstracts cutoff.

# abstracts	True positives	False positives	Total predictions	False negatives
>=1	225	2236	2461	1497
>=2	130	721	851	1592
>=3	89	362	451	1633
>=4	78	254	332	1644
>=5	61	193	254	1661
>=6	52	108	160	1670
>=7	44	81	125	1678
>=8	39	63	102	1683
>=9	33	50	83	1689
>=10	30	42	72	1692
>=11	28	35	63	1694
>=12	27	31	58	1695

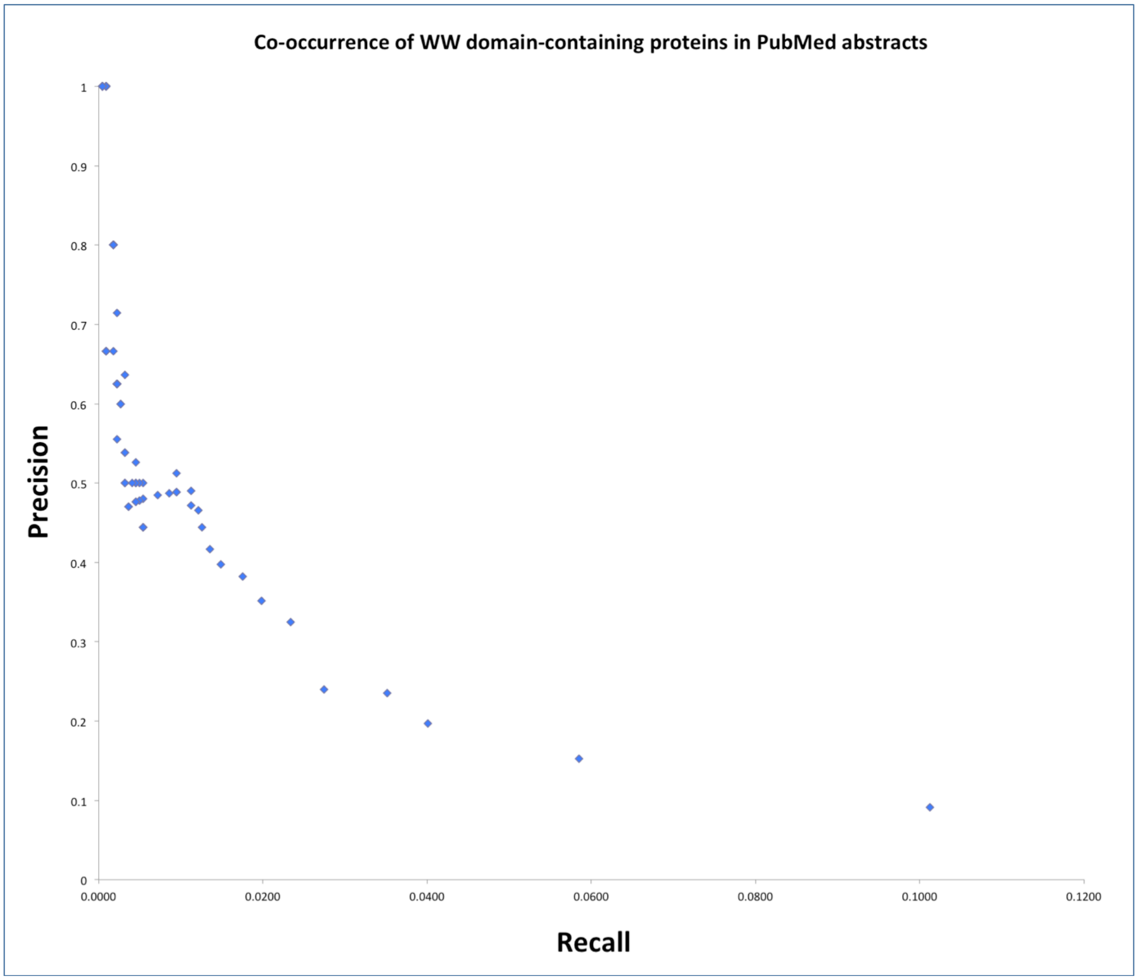
Table 2.3.7: WW domain-containing protein interaction prediction results based on co-occurrence of protein names in PubMed abstracts (2012)

# abstracts cutoff	Recall	Precision
>=1	0.1307	0.0914
>=2	0.0868	0.1528
>=3	0.0651	0.1973
>=4	0.0610	0.2349
>=5	0.0508	0.2402
>=6	0.0457	0.3250
>=7	0.0405	0.3520
>=8	0.0374	0.3824
>=9	0.0329	0.3976
>=10	0.0309	0.4167
>=11	0.0298	0.4444
>=12	0.0296	0.4655

As evident from this figure, precision drops as the abstract cutoff decreases. Recall is also low at high precision.

The resulting precision-recall graph is shown in **Fig. 2.3.8**.

Figure 2.3.8: Precision-recall graph of WW domain-containing PPI predictions by text mining based on co-occurrence of protein names in PubMed abstracts



2.4 – DISCUSSION

The main problem with text mining by gene name co-occurrence, besides long execution time, is **ambiguity**. For gene names that match a dictionary word, such as *ITCH*, PubMed search returns all abstracts that contain this word and its derivatives (‘itchy skin’, ‘itching’, etc.), not necessarily the protein name. The same applies to gene names consisting of one or two characters, such as ‘T’ or ‘TH’, which form parts of English words, and gene names equivalent to disease name abbreviations, such as ‘MS’. The number of PubMed abstracts returned for these genes exceeds several thousand, resulting in a high number of false positives, which, in turn, lead to low recall.

2.5 – CONCLUSIONS

Text mining by gene name co-occurrence in PubMed abstracts is successful in predicting interacting protein pairs with precision rate between approximately 0.1 and 0.5, and recall rate between approximately 0.03 and 0.1.

CHAPTER 3: MeSH TERM INDEXING

To resolve the issue of ambiguity mentioned in 2.4, I needed to refine the search to limit the number of abstracts returned by the search to abstracts that specifically talk about this gene name. For this, I used the **Me**dical **S**ubject **H**eading (**MeSH**) **term indexing**.

The **Me**dical **S**ubject **H**eading (**MeSH**) database is a controlled vocabulary of the U.S. National Library of Medicine that uniformly indexes biomedical literature (NIH: U.S. National Library of Medicine 2012). The MeSH vocabulary includes four main types of terms: Headings (descriptors), Subheadings (qualifiers), Supplementary Concepts, and Publication Types (NIH: U.S. National Library of Medicine 2012). These terms characterize different aspects of the published MeSH records and are classified as Descriptors, Qualifiers, or Supplementary Concept Records (SCRs) (NIH: National Library of Medicine 2014). For human protein names, the MeSH terms are in the form "***official_gene_symbol* protein, human**" (e.g. '**BAG3 protein, human**') and are indexed in PubMed abstracts as either 'Supplementary Concepts' or 'Entry terms'.

MeSH term indexing facilitates searching PubMed by retrieving only publications that discuss the search term substantively. A publication that simply mentions a concept but does not discuss it in detail is not indexed with a MeSH term for this concept and will not be returned by the search. Hence, using the MeSH Supplementary Concept indexing of protein names to retrieve only articles that specifically discuss these proteins and their interacting partners is expected to increase search precision considerably, compared to only searching for simple gene name mentions.

3.1 – MATERIALS AND METHODS

3.1.1. Datasets:

The text mining process started with the following datasets:

- A list of all approved official HGNC symbols for every protein-coding gene in the human genome (**32717** in total) in text format – **Set A**

- A list of **50** gene symbols of WW domain-containing proteins, downloaded from **Biomart** (Smedley et al. 2009, Zhang et al. 2011) in text format – **Set B** **(Appendix B)**

- A list of **MeSH** terms for all genes from **Set A** and **Set B**. These terms may be retrieved in batch from the MeSH database using Python EUtils package as follows:

```
handle = Entrez.esearch(db="mesh", term=prot, rettype='xml',
retmax=ret_max)
```

A unique MeSH supplementary concept term was found for 44 of the 50 WW domain-containing proteins. Of the remaining 6 proteins, 5 have not been indexed for MeSH, and the MeSH supplementary concept for DRP2 is not in the format ‘DRP2 protein, human’ and was excluded from the search.

The returned abstract set corresponds to the results of a generalized manual search of PubMed using the MeSH term as a search keyword. Alternatively, the MeSH terms may be downloaded manually in XML format from the MeSH repository. Then Python and/or UNIX commands would be used to extract the protein names and their corresponding MeSH terms from the downloaded file, in the form “gene_name protein, human” (e. g. “A1CF protein, human”, which corresponds to the ‘A1CF’ gene symbol). However, the script’s execution time using MEDLINE is many times faster than retrieving the records as XML or plain text.

The retrieved MeSH terms were stored in a MySQL database.

- A set of all PubMed abstracts for every WW-containing protein in **Set B**, excluding DMD and ITCH, since they were not included in part 1 of the analysis. The abstracts were retrieved using the built-in ‘*esearch*’ and ‘*efetch*’ methods of the Entrez Programming Utilities Python package (**EUtils**) (National Center for Biotechnology Information 2008).

  Alternatively, the abstracts can be downloaded manually from the NLM MeSH website: http://www.nlm.nih.gov/mesh/filelist.html. Either of these techniques may be implemented for use in the future in an automated prediction pipeline. Approach a) requires no human interaction, whereas approach b) requires human effort.

The retrieved abstracts with their PMIDs were stored in a MySQL database.

- A list of **1722** interacting protein pairs, in text format, downloaded from iRefWeb in full MITAB format (same as in Chapter 2) using the script from **Appendix E**.

Computational resources are the same as in Chapter 2.

3.1.2. Text mining:

Once the datasets were downloaded, Python scripts were written and executed to find the MeSH terms for protein names from Set A from the downloaded PubMed abstracts. Proteins whose corresponding MeSH terms were indexed in the same abstract were predicted to interact. The interacting pair, along with the PubMed publication IDs (PMIDs) of the abstracts in which the interaction was found, were recorded in a tab-delimited file. Interactions were then grouped by number of abstracts where each interaction was found (similar to Chapter 2), precision and recall were computed, and a precision-recall graph was constructed.

Scripts:

- **mesh_search.py:** **(Appendix I)**

Contains code to download the MeSH term for every human protein-coding gene symbol. The script accesses the MeSH repository remotely. Retrieved MeSH term records were stored in a MySQL database

- **eutils_search.py:**

This script does two things:

a) Search PubMed remotely for all abstracts pertinent to each WW-containing protein from Set B.

b) In every abstract, identify all MeSH terms indexed in it, both for WW-containing proteins and any other proteins from the set of HUGO gene symbols. All MeSH terms were recorded in the database as predicted interacting partners, along with the PMID of the abstract where they were found.

The resulting interaction dataset was stored in the database.

Table 3.1.1: A sample MySQL table storing mapping between HGNC symbols and their corresponding MeSH terms:

MeSH Term	Gene Symbol
UTRN protein, human	UTRN
DMD protein, human	DMD
ARHGAP27 protein, human	ARHGAP27
GAS7 protein, human	GAS7
ITCH protein, human	ITCH

Table 3.1.2: A sample MySQL table storing predicted interactions for MeSH terms indexed in the same abstract.

Source MeSH Term	Interacting Partner	PMID of abstracts where interactions were found
GAS7 protein, human	RUNX2 protein, human	21452305
ITCH protein, human	LATS protein, human	21383157, 21212414
SMURF1 protein, human	SMURF2 protein, human	22351504,20937913, 20484049

Table 3.1.3: WW domain-containing protein interaction prediction results based on MeSH term indexing in PubMed abstracts (2013)

# Abstracts	True positives	False positives	Total predictions	False negatives
>=1	85	556	641	1637
>=2	25	70	95	1697
>=3	7	20	27	1715
>=4	3	8	11	1719
>=5	2	3	5	1720
>=6	1	1	2	1721
>=7	1	1	2	1721
>=8	1	1	2	1721
>=9	1	0	1	1721
>=10	1	0	1	1721
>=11	1	0	1	1721
>=12	1	0	1	1721

Table 3.1.4: Precision and recall computed for results in Fig. 3.1.3

# abstracts cutoff	Recall	Precision
>=1	0.0494	0.1326
>=2	0.0145	0.2632
>=3	0.0041	0.2593
>=4	0.0017	0.2727

>=5	0.0012	0.4000
>=6	0.0006	0.5000
>=7	0.0006	0.5000
>=8	0.0006	0.5000
>=9	0.0006	1.0000
>=10	0.0006	1.0000
>=11	0.0006	1.0000
>=12	0.0006	1.0000

Figure 3.1.5: Precision-recall graph of WW domain-containing PPI predictions by text mining based on MeSH terms indexed in PubMed abstracts

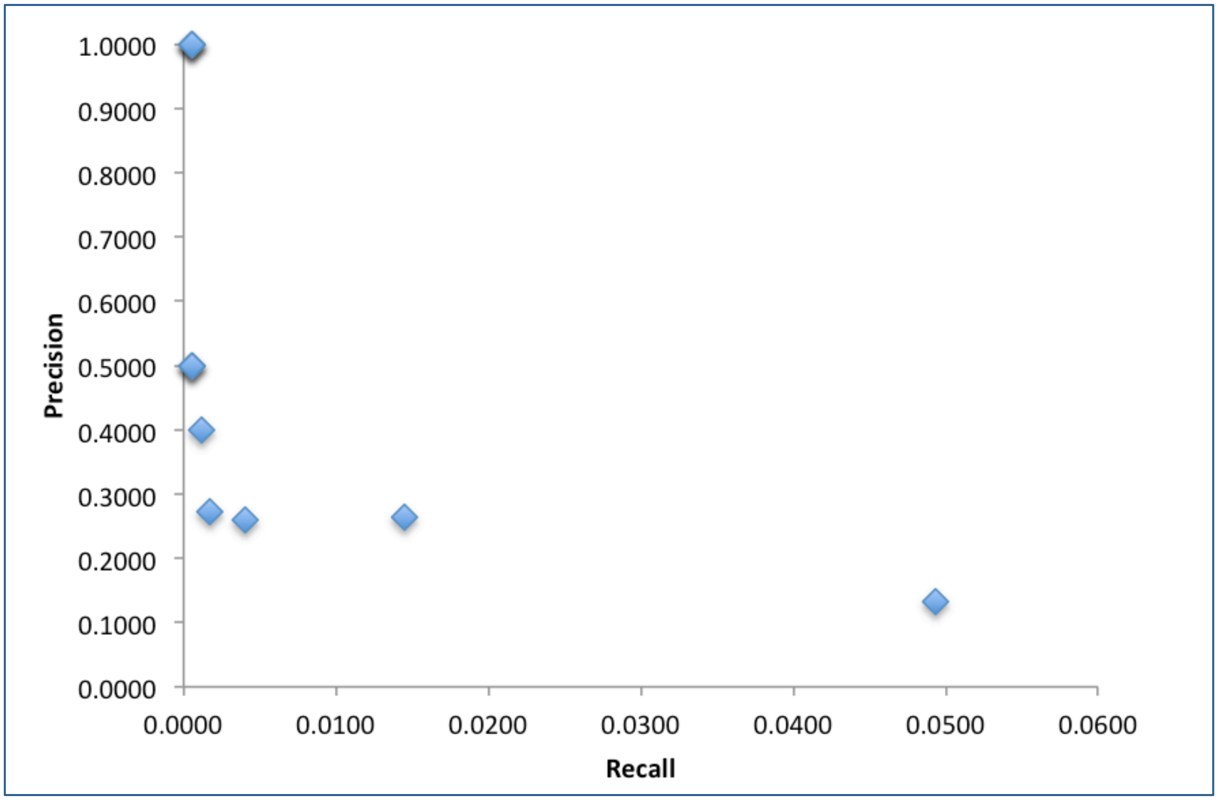
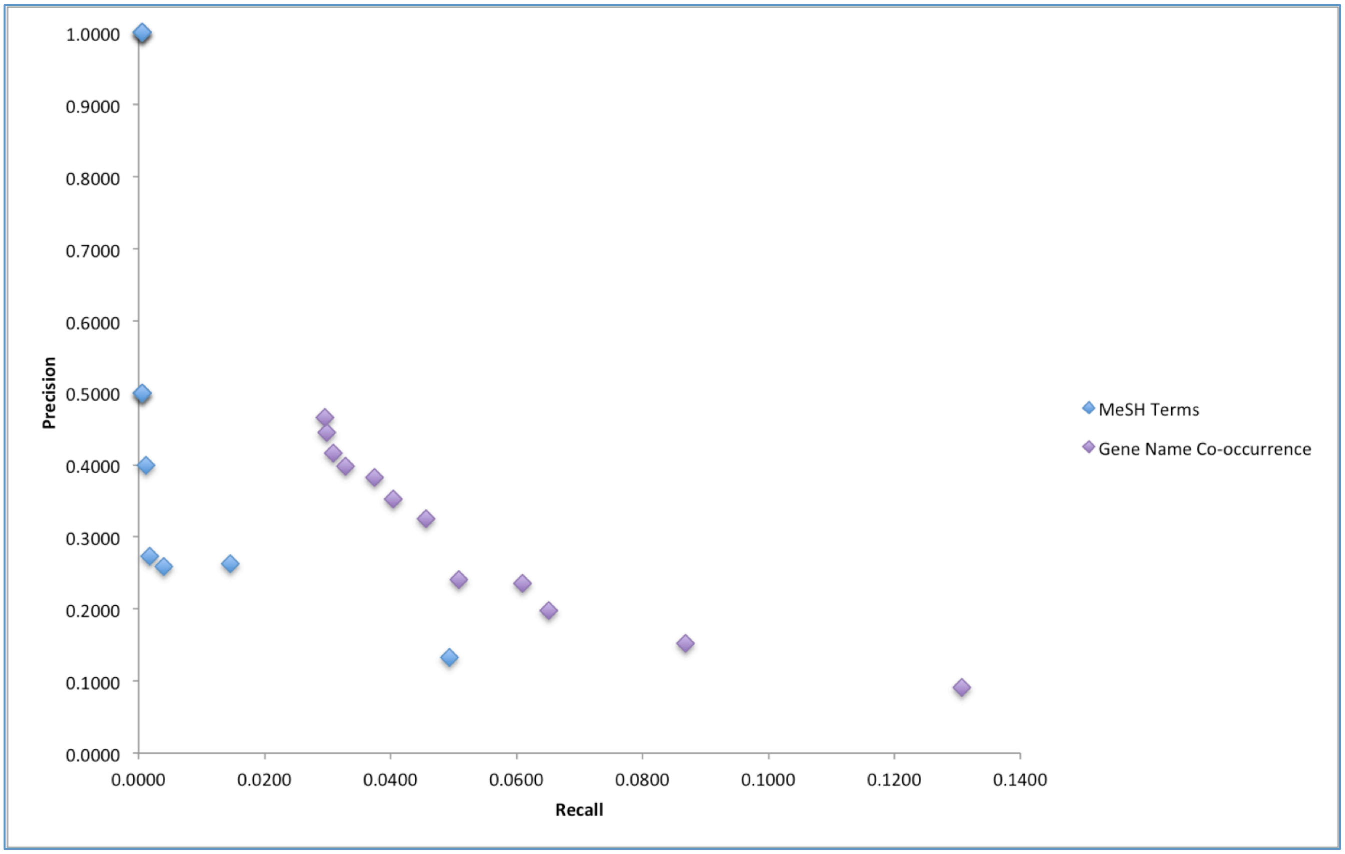


Table 3.1.6: Comparison of text mining prediction results based on gene name co-occurrence in abstracts and results based on MeSH term indexing

	Gene name co-occurrence		MeSH terms	
Abstracts	**Recall**	**Precision**	**Recall**	**Precision**
>=1	0.1307	0.0914	0.0494	0.1326
>=2	0.0868	0.1528	0.0145	0.2632
>=3	0.0651	0.1973	0.0041	0.2593
>=4	0.0610	0.2349	0.0017	0.2727
>=5	0.0508	0.2402	0.0012	0.4000
>=6	0.0457	0.3250	0.0006	0.5000
>=7	0.0405	0.3520	0.0006	0.5000
>=8	0.0374	0.3824	0.0006	0.5000
>=9	0.0329	0.3976	0.0006	1.0000
>=10	0.0309	0.4167	0.0006	1.0000
>=11	0.0298	0.4444	0.0006	1.0000
>=12	0.0296	0.4655	0.0006	1.0000

A comparison of both curves is illustrated in Fig. 3.1.7.

Figure 3.1.7: Comparison of text mining results based on gene name co-occurrence to results based on MeSH term indexing.



3.1.3 – DISCUSSION

The main disadvantage of using MeSH indexing for PPI prediction is the lack of indexing for many proteins that are not a major topic of an article. If a protein has not been indexed for MeSH, it would not be considered an interacting partner, even though it might be a TP.

3.1.4 – CONCLUSIONS

Text mining results based on MeSH term indexing show much greater precision than results based on gene name co-occurrence, as expected, due to stringent definition of MeSH terms and publication indexing in PubMed. However, the recall is much lower than recall using gene name co-occurrence. Many proteins that interact in real life and are co-mentioned in abstracts but are not major topics of publications are not indexed for MeSH. For this reason, a script that relies on MeSH term indexing to predict protein interactions returns a high number of false negatives, i.e. predictions are simply missed. MeSH terms can, therefore, be used for PubMed text mining in cases where high precision is required; however, due to the low recall, it is not worthwhile to use MeSH terms for large-scale text mining.

CHAPTER 4. CO-OCCURRENCE WITH CONTROLLED VOCABULARY

Since MeSH term search yielded poor recall, the final step of the analysis reverts to gene name co-occurrence. To improve precision by reducing the number of false positives and predict interactions that specifically involve the WW domain, a controlled vocabulary has been applied to identify specific interactions mediated by the WW domain.

The controlled vocabulary terms have been selected manually based on known WW PPI abstracts, to reflect WW domain-mediated interactions to the domain's binding motif. The list includes the following terms:

- ***proline-rich***
- ***protein-protein interaction***
- ***WW domain***
- ***proline residues***
- ***PPxY***
- ***LPxY***
- ***protein interaction module***
- ***WW domain-binding***
- ***WW-binding***

4.1. MATERIALS AND METHODS

4.1.1. Datasets:

The text mining process started with the following datasets:

- A list of all approved official HGNC symbols for every protein-coding gene in the human genome (**32717** in total) in text format – **Set A**

- A list of **50** gene symbols of WW domain-containing proteins, downloaded from **Biomart** (Smedley et al. 2009, Zhang et al. 2011) in text format – **Set B** **(Appendix B)**

- A list of controlled vocabulary terms – **Set C**

Set C: Controlled vocabulary terms

- ***proline-rich***
- ***protein-protein interaction***
- ***WW domain***
- ***proline residues***
- ***PPxY***
- ***LPxY***
- ***protein interaction module***
- ***WW domain-binding***
- ***WW-binding***

- A set of all PubMed abstracts for every WW-containing protein in **Set B**, excluding DMD and ITCH, since they were not included in part 1 of the analysis. The abstracts were retrieved using the built-in '*esearch*' and '*efetch*' methods of the Entrez Programming Utilities Python package (**EUtils**) (National Center for Biotechnology Information (US) 2008) –and stored in a custom MySQL database.

Alternatively, the abstracts can be downloaded manually from the NLM MeSH website: http://www.nlm.nih.gov/mesh/filelist.html. Either of these techniques may be implemented for use in the future in an automated prediction pipeline. Approach a) requires no human interaction, whereas approach b) requires human effort.

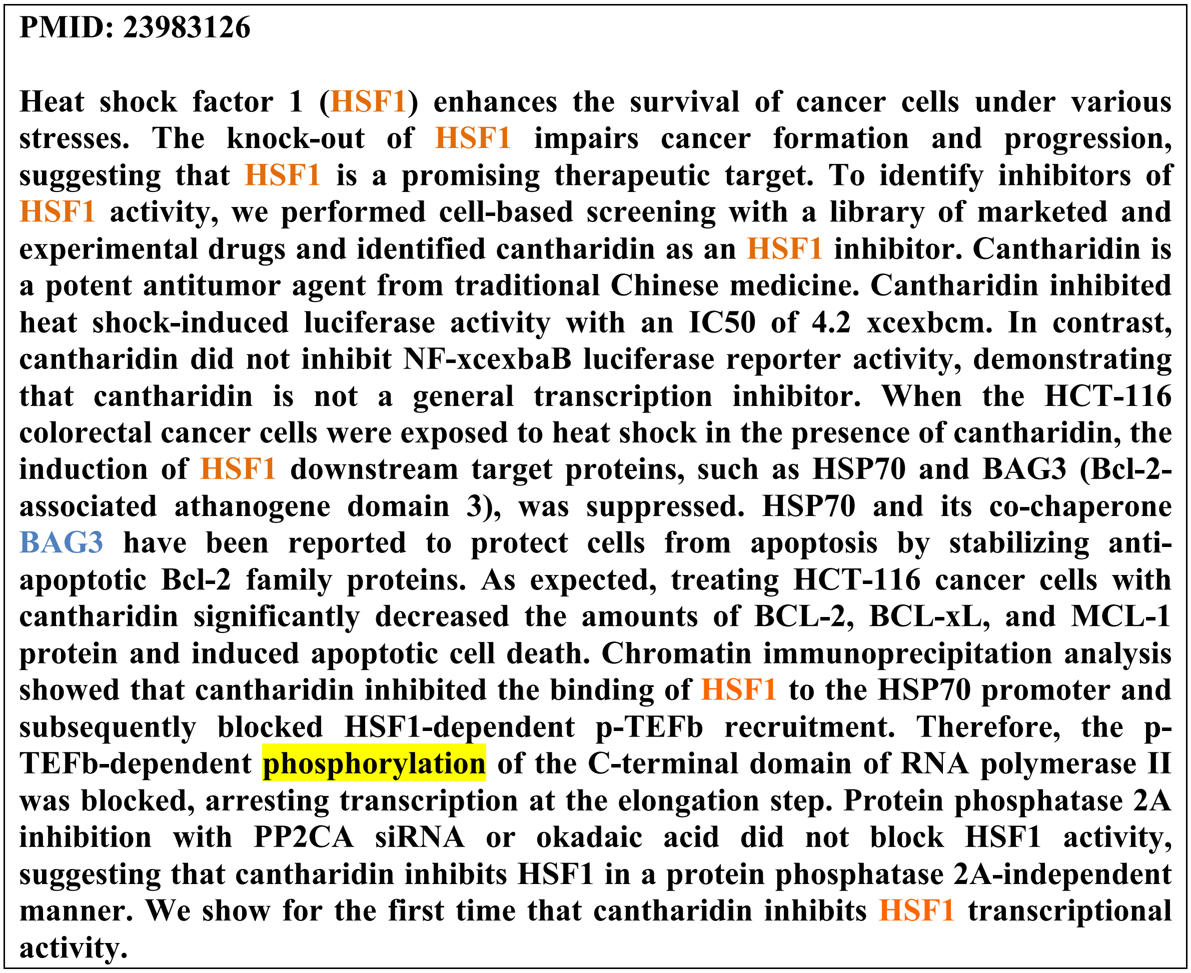
Computational resources (UNIX, Python, MySQL) are the same as in Chapters 2 and 3.

4.1.2. Text mining:

For each controlled vocabulary term, abstracts that contain this term, were selected for analysis. Remaining abstracts were discarded.

An example of such an abstract is shown in Fig. 4.1.1

Figure 4.1.1: An abstract that contains controlled vocabulary term 'PHOSPHORYLATION'. Co-occurring protein names HSF1 and BAG3 are highlighted. These proteins were predicted to interact.

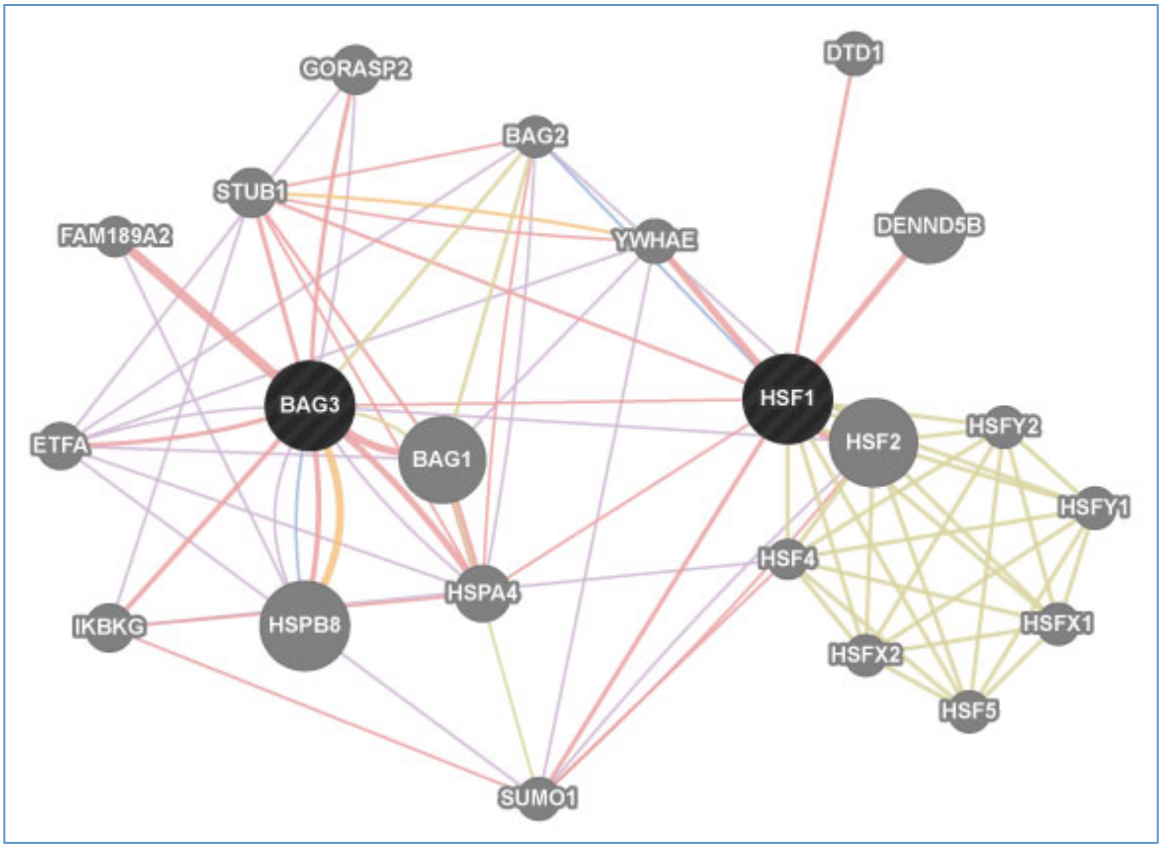


PMID: 23983126

Heat shock factor 1 (HSF1) enhances the survival of cancer cells under various stresses. The knock-out of HSF1 impairs cancer formation and progression, suggesting that HSF1 is a promising therapeutic target. To identify inhibitors of HSF1 activity, we performed cell-based screening with a library of marketed and experimental drugs and identified cantharidin as an HSF1 inhibitor. Cantharidin is a potent antitumor agent from traditional Chinese medicine. Cantharidin inhibited heat shock-induced luciferase activity with an IC50 of 4.2 xcexbcm. In contrast, cantharidin did not inhibit NF-xcexbaB luciferase reporter activity, demonstrating that cantharidin is not a general transcription inhibitor. When the HCT-116 colorectal cancer cells were exposed to heat shock in the presence of cantharidin, the induction of HSF1 downstream target proteins, such as HSP70 and BAG3 (Bcl-2-associated athanogene domain 3), was suppressed. HSP70 and its co-chaperone BAG3 have been reported to protect cells from apoptosis by stabilizing anti-apoptotic Bcl-2 family proteins. As expected, treating HCT-116 cancer cells with cantharidin significantly decreased the amounts of BCL-2, BCL-xL, and MCL-1 protein and induced apoptotic cell death. Chromatin immunoprecipitation analysis showed that cantharidin inhibited the binding of HSF1 to the HSP70 promoter and subsequently blocked HSF1-dependent p-TEFb recruitment. Therefore, the p-TEFb-dependent phosphorylation of the C-terminal domain of RNA polymerase II was blocked, arresting transcription at the elongation step. Protein phosphatase 2A inhibition with PP2CA siRNA or okadaic acid did not block HSF1 activity, suggesting that cantharidin inhibits HSF1 in a protein phosphatase 2A-independent manner. We show for the first time that cantharidin inhibits HSF1 transcriptional activity.

Proteins HSF1 and BAG3, which co-occur in the referenced article, also interact in vivo, as shown by the Cytoscape (Shannon et al. 2003) diagram in Fig. 4.1.2.

Figure 4.1.2: A Cytoscape (Shannon et al. 2003) diagram illustrating the interaction between BAG3 and HSF1 proteins



Interacting protein pairs and the PMIDs of the abstract where they co-occur were stored in a MySQL database. Again, co-occurrences were grouped by the number of abstracts, and a precision-recall graph was constructed.

The script was first run for each controlled vocabulary (CV) term; then a combined prediction script was run for all the CV terms.

Prediction statistics using co-occurrence with CV terms and comparison to previous text mining results are shown in **Fig. 4.1.4 – 4.1.8**.

Table 4.1.3: The number of TPs and FPs for each individual CV term

Term	Total predictions	TP	FP
LPxY	1	0	1
PPxY	93	23	70
Proline residues	8	1	7
Proline-rich	45	9	36
Protein interaction	56	15	41
PXXP	1	0	1
WW-binding	9	2	7
WW domain	321	74	247
WW domain-binding	4	2	2
Total:	539		

Figure 4.1.4: Precision-recall table and graph for CV term 'protein interaction'

>= abstracts	Recall	Precision
>=1	0.0065	0.2245
>=2	0.0012	0.4000
>=4	0.0006	1.0000

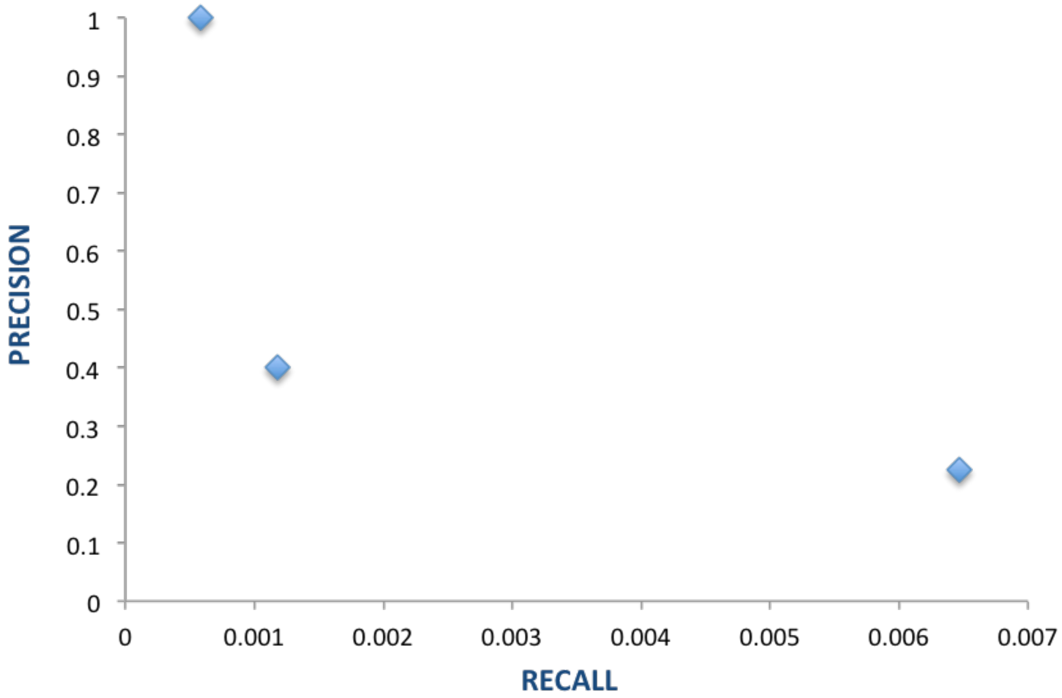


Figure 4.1.5: Precision-recall table and graph for CV term 'WW domain'

abstract cutoff	Recall	Precision
>=1	0.0306	0.2271
>=2	0.0094	0.2807
>=3	0.0029	0.2500
>=4	0.00059	0.1250

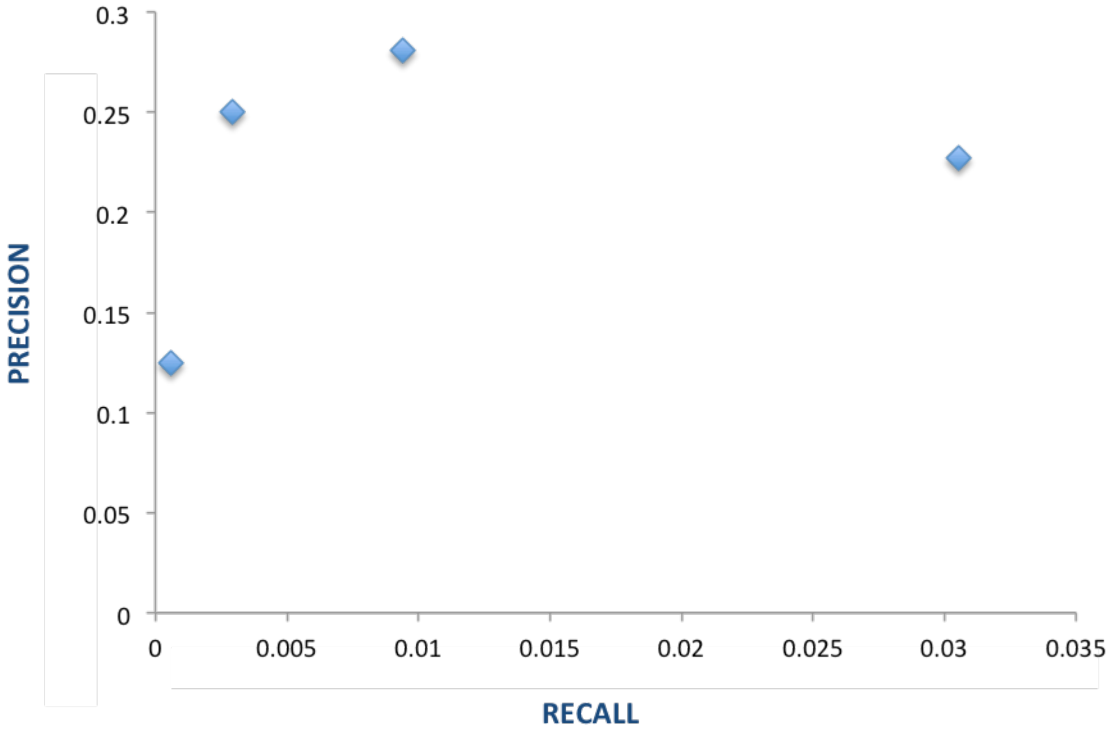


Figure 4.1.6: Precision-recall table and graph for CV term 'PPXY'

>= abstracts	Recall	Precision
>=1	0.0106	0.2903
>=2	0.0024	0.2857
>=3	0.0006	0.2000
>=4	0	0

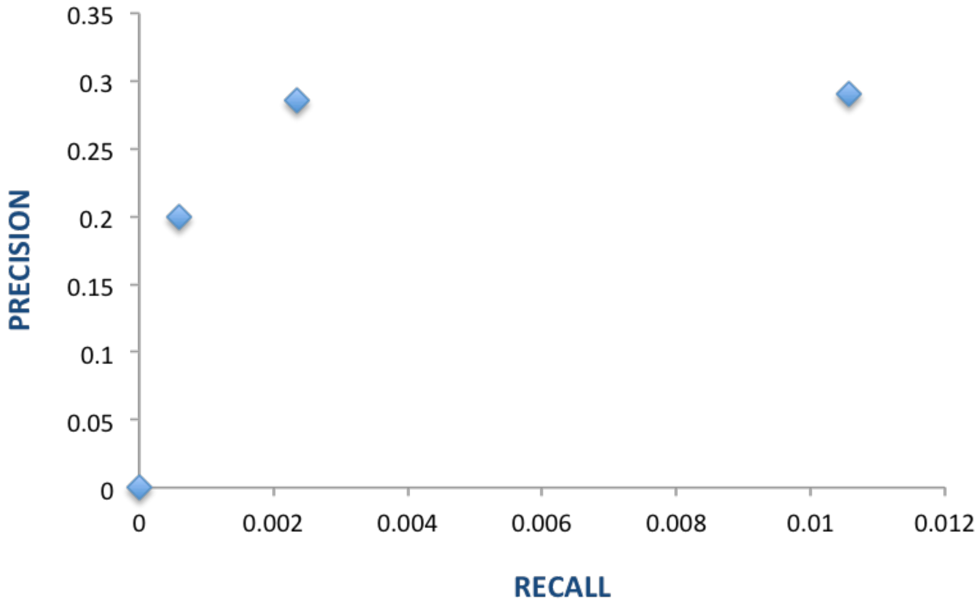
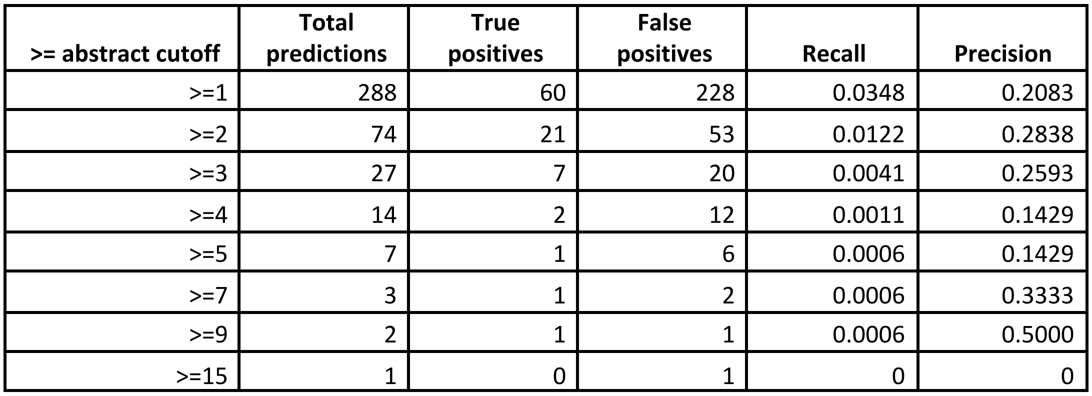


Figure 4.1.7: **Precision and recall values for all CV terms combined**



>= abstract cutoff	Total predictions	True positives	False positives	Recall	Precision
>=1	288	60	228	0.0348	0.2083
>=2	74	21	53	0.0122	0.2838
>=3	27	7	20	0.0041	0.2593
>=4	14	2	12	0.0011	0.1429
>=5	7	1	6	0.0006	0.1429
>=7	3	1	2	0.0006	0.3333
>=9	2	1	1	0.0006	0.5000
>=15	1	0	1	0	0

Figure 4.1.8: Precision-recall graph for gene name co-occurrence with all CV terms combined

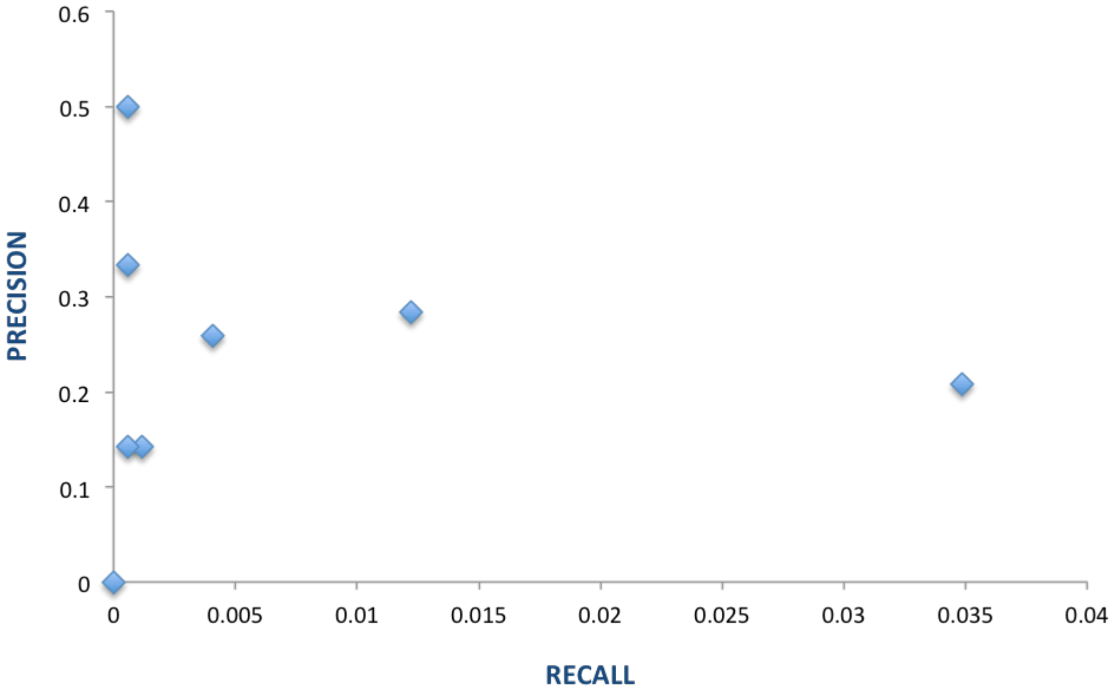
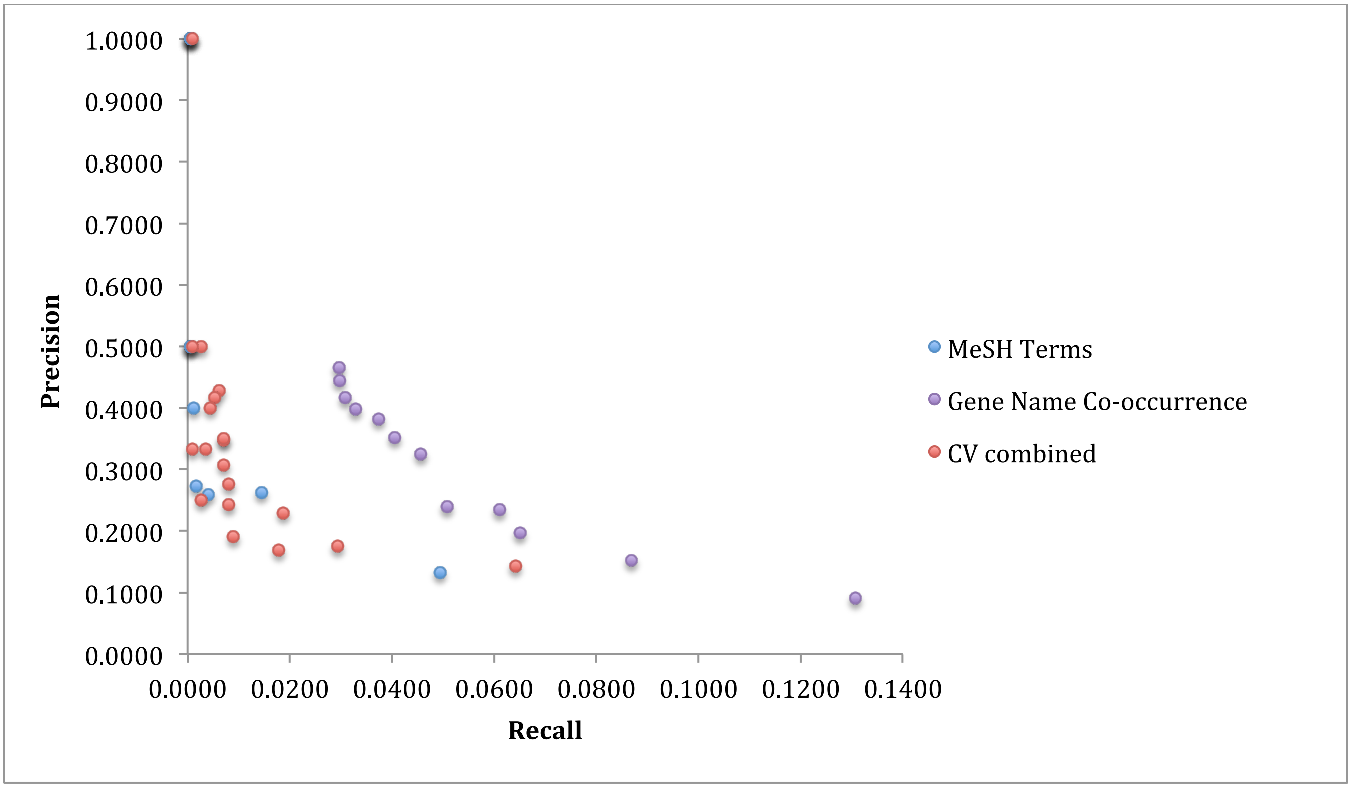


Figure 4.1.9: Comparison of precision and recall values for gene name co-occurrence with and without all CV terms combined

	Gene name co-occurrence without CV		Gene name co-occurrence with CV	
Abstracts	Recall	Precision	Recall	Precision
>=1	0.1307	0.0914	0.0348	0.2083
>=2	0.0868	0.1528	0.0122	0.2838
>=3	0.0651	0.1973	0.0041	0.2593
>=4	0.0610	0.2349	0.0012	0.1429
>=5	0.0508	0.2402	0.0006	0.1429
>=6	0.0457	0.3250	0.0006	0.1429
>=7	0.0405	0.3520	0.0006	0.3333
>=8	0.0374	0.3824	0.0006	0.3333
>=9	0.0329	0.3976	0.0006	0.5
>=10	0.0309	0.4167	0	0
>=11	0.0298	0.4444	0	0
>=12	0.0296	0.4655	0	0

Figure 4.1.10: Comparison of all text mining results – based on gene name co-occurrence without CV, based on MeSH terms, and based on gene name co-occurrence with CV



4.2. DISCUSSION

As expected, text mining based on co-occurrence with CV yields higher precision and higher recall than co-occurrence without CV. It also yields higher precision than MeSH term-based mining. It shows lower recall than MeSH term-based mining, which can be explained by two factors. The first is the stringency of CV selection – like MeSH term definition, the CV algorithm may discard relevant abstracts that contain information on PPIs but not the CV terms.

As outlined in section 2.4, it would be helpful to rerun the prediction script with a new version of iRefIndex and ensure that the publications examined by the script are not newer than the latest version of the benchmark. This step may help identify more TPs and fewer FPs.

4.3. CONCLUSIONS

Based on the precision-recall graph for gene name co-occurrence with CV, it may be concluded that gene name co-occurrence is a text mining algorithm that can be used as a novel evidence source, for PPIs that involve the WW domain and for other PPIs.

FUTURE DIRECTIONS

It would be helpful to consider gene name aliases. Since many proteins are listed in HGNC under multiple names, or their names have changed over time, and the HGNC symbol differs from the name in the publication, gene names may be missed by the co-occurrence script. Moreover, different databases use different identifiers for the same gene (e.g. NCBI and Ensembl), and publication authors may use these identifiers interchangeably. Therefore, it would be helpful to have a method to detect multiple gene names and different database identifiers.

Also, some of the abstracts have been published after the release of iRefIndex v.4.1; therefore, PPIs outlined in these abstracts would be identified as FPs when they are actually TPs. Rerunning the script with a new version of iRefIndex, as well as ensuring that the publications examined by the script are not newer than the latest version of the benchmark may help identify more TPs and fewer FPs.

The pipeline needs to be expanded to other domains, besides WW-containing proteins, and updated with the latest version of iRefIndex as a benchmark. Abstract selection needs to reflect the dates of the publications to correlate with the benchmark and not include abstracts published after the benchmark release.

The controlled vocabulary needs to be expanded and adapted to more natural language processing (e.g. 'does not interact', 'but not' or 'except') [**Fig. 5.1**, NLP terms highlighted in red]. Furthermore, it would be helpful to include the distance between terms and assign weights to interactions based on how far protein names are from each other in the abstract.

Figure 5.1: An abstract describing the interaction of Smurf2 and Smad1 proteins *in vivo* (Lin et al. 2000)

Smads are important intracellular signaling effectors for transforming growth factor-beta (TGF-beta) and related factors. Proper TGF-beta signaling requires precise control of Smad functions. In this study, we have identified a novel HECT class ubiquitin E3 ligase, designated Smurf2, that negatively regulates Smad2 signaling. In both yeast two-hybrid and in vitro binding assays, we found that Smurf2 could interact with receptor-activated Smads (R-Smads), including Smad1, Smad2, and Smad3 but not Smad4. Ectopic expression of Smurf2 was sufficient to reduce the steady-state levels of Smad1 and Smad2 but not Smad3 or Smad4. Significantly, Smurf2 displayed preference to Smad2 as its target for degradation. Furthermore, Smurf2 exhibited higher binding affinity to activated Smad2 upon TGF-beta stimulation. The ability of Smurf2 to promote Smad2 destruction required the HECT catalytic activity of Smurf2 and depended on the proteasome-dependent pathway. Consistent with these results, Smurf2 potently reduced the transcriptional activity of Smad2. These data suggest that a ubiquitin/proteasome-dependent mechanism is important for proper regulation of TGF-beta signaling.

J Biol Chem. 2000 Nov 24;275(47):36818-22.
Smurf2 is a ubiquitin E3 ligase mediating proteasome-dependent degradation of Smad2 in transforming growth factor-beta signaling.
Lin X[1], Liang M, Feng XH.

BIBLIOGRAPHY

Amartely H, Iosub-Amir A, Friedler A. "Identifying protein-protein interaction sites using peptide arrays." *J Vis Exp.*, 2014.

Auerbach D, Stagljar I. "Yeast Two-Hybrid Protein–Protein Interaction Networks."

Berggård T, Linse S, James P. "Methods for the detection and analysis of protein-protein interactions." *Proteomics*, August 2007: 2833-42.

Bork P, Sudol M. "The WW domain: a signalling site in dystrophin?" *Trends Biochem Sci.*, 1994.

Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. "Yeast Two-Hybrid, a Powerful Tool for Systems Biology." *Int J Mol Sci.*, 2009.

Chen XW, Jeong JC. "Sequence-based prediction of protein interaction sites with an integrative method." *Bioinformatics*, 2009.

Cohen KB, Hunter L. "Getting Started in Text Mining." *PLoS Comput Biol.*, Jan 2008.

Darnell, JE Jr. "STATs and gene regulation." *Science*, Sep 1997.

Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. "Crystal Structures of a Complexed and Peptide-Free Membrane Protein–Binding Domain: Molecular Basis of Peptide Recognition by PDZ." *Cell*, 1996.

Fields S, Song O. "A novel genetic system to detect protein-protein interactions." *Nature*, 1989: 245-6.

Fukasawa H, Yamamoto T, Togawa A, Ohashi N, Fujigaki Y, Oda T, Uchida C, Kitagawa K, Hattori T, Suzuki S, Kitagawa M, Hishida A. "Down-regulation of Smad7 expression by ubiquitin-dependent degradation contributes to renal fibrosis in obstructive nephropathy in mice." *Proc Natl Acad Sci U S A.* , Jun 2004.

George, Susan R., O'Dowd, Brian F. *G Protein-Coupled Receptor-Protein Interactions.* Toronto: John Wiley & Sons, Inc., 2005.

Gonzalez MW, Kann MG. "Chapter 4: Protein Interactions and Disease." *PLOS Computational Biology*, 2012.

Gonzalez, Kann. "Chapter 4: Protein Interactions and Disease." *PLOS Computational Biology*, 2012.

Hall, Randy A. "CO-IMMUNOPRECIPITATION AS A STRATEGY TO EVALUATE RECEPTOR–RECEPTOR OR RECEPTOR–PROTEIN INTERACTIONS." In *G Protein-Coupled Receptor--Protein Interactions*, by Susan R., O'Dowd, Brian F. George, 165. Toronto: Wiley & Sibs, 2004.

Hill T., Lewicki P. *Statistics: Methods and Applications.* Tulsa, OK: Dell, 2007.

Hosur, R. "Structure-based algorithms for protein-protein interaction prediction." Boston, MA: DSpace@MIT http://dspace.mit.edu/handle/1721.1/75843, June 2012.

Hu H, Columbus J, Zhang Y, Wu D, Lian L, Yang S, Goodwin J, Luczak C, Carter M, Chen L, James M, Davis R, Sudol M, Rodwell J, Herrero JJ. "A map of WW domain family interactions." *Proteomics*, 2004.

HUPO Proteomics Standards Initiative. *Molecular Interaction XML Format Documentation.* http://psidev.sourceforge.net/molecular_interactions/xml/doc/user/ .

Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CS, Yu J, Hersi K, Raaijmakers J, Gish G, Mbamalu G, Taylor L, Yeung B, Vassilovski G, Amin M, Chen F, Matskova L, Winberg G, Ernberg I, Linding R, O'donnell P, Starostine A, Keller W, Metalnikov P, Stark C, Pawson T. "WW domains provide a platform for the assembly of multiprotein networks." 2005.

Isono E, Schwechheimer C. "Co-immunoprecipitation and protein blots." *Methods Mol Biol.*, 2010.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *Proc Natl Acad Sci U S A.*, 2001: 4569-74.

Jain S, Bader G. D. "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology." *BMC Bioinformatics*, 2010.

Jenssen TK, Laegreid A, Komorowski J, Hovig E. "A literature network of human genes for high-throughput analysis of gene expression." *Nat Genet.*, May 2001.

Katz C, Levy-Beladev L, Rotem-Bamberger S, Rito T, Rüdiger SG, Friedler A. "Studying protein-protein interactions using peptide arrays." *Chem Soc Rev.*, 2011.

Kay BK, Castagnoli L. "Mapping protein-protein interactions with phage-displayed combinatorial peptide libraries. ." *Curr Protoc Cell Biol.* , 2003.

Kennedy, MB. "Origin of PDZ (DHR, GLGF) domains." *Trends Biochem Sci.*, Sep 1995.

Kerpedjiev P, Frellsen J, Lindgreen S, Krogh A. "Adaptable probabilistic mapping of short reads using position specific scoring matrices." *BMC Bioinformatics*, 2014.

Kim, E., Sheng M. "PDZ domain proteins of synapses." *Nat Rev Neurosci*, 2004.

Kodaka M, Hata Y. "The mammalian Hippo pathway: regulation and function of YAP1 and TAZ ." *Cellular and Molecular Life Sciences*, January 2015.

Kokoszka ME, Kay BK. "Mapping protein-protein interactions with phage-displayed combinatorial peptide libraries and alanine scanning. ." *Methods Mol Biol.*, 2015.

Krallinger M, Valencia A and Hirschman L. "Linking genes to literature: text mining, information extraction, and retrieval applications for biology." *Genome Biology*, 2008.

Lee JM, Sonnhammer EL. "Genomic Gene Clustering Analysis of Pathways in Eukaryotes." *Genome Res.*, 2003: 875-82.

Leser U, Hakenberg J. "What makes a gene name? Named entity recognition in the biomedical literature." *Brief Bioinform.* , 2005.

Lin J, Wilbur WJ. "PubMed related articles: a probabilistic topic-based model for content similarity." *BMC Bioinformatics*, Oct 2007.

Lin X, Liang M, Feng XH. "Smurf2 is a ubiquitin E3 ligase mediating proteasome-dependent degradation of Smad2 in transforming growth factor-beta signaling." *J Biol Chem.* , Nov 2000.

Liu X, Liu B, Huang Z, Shi T, Chen Y, Zhang J. "SPPS: A Sequence-Based Method for Predicting Probability of Protein-Protein Interaction Partners." *PLoS One*, 2012.

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. "Cytoscape Web: an interactive web-based network browser." *Bioinformatics* 26, no. 18 (September 2010): 2347-8.

Mayer, BJ. "SH3 domains: complexity in moderation." *J Cell Sci.*, 2001.

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function." *Genome Biol.*, 2008.

Mostarda, S., D. Gfeller, et al. "Beyond the binding site: the role of the beta(2)-beta(3) loop and extra-domain structures in PDZ domains." *PLoS Comput Biol*, 2012.

Musacchio A, Gibson T, Lehto VP, Saraste M. "SH3--an abundant protein domain in search of a function." *FEBS Lett.* , 1992.

National Center for Biotechnology Information (US). "Entrez Programming Utilities Help." *Entrez Programming Utilities Help.* December 12, 2008. http://www.ncbi.nlm.nih.gov/books/NBK25501/ (accessed 2015).

Nguyen JT, Turck CW, Cohen FE, Zuckermann RN, Lim WA. "Exploiting the basis of proline recognition by SH3 and WW domains: design of N-substituted inhibitors." *Science*, 1998.

NIH: National Library of Medicine. *Medical Subject Headings.* August 06, 2014. https://www.nlm.nih.gov/mesh/intro_record_types.html (accessed 2015).

NIH: U.S. National Library of Medicine. *Medical Subject Headings (MeSH®) in MEDLINE®/PubMed®: A Tutorial.* October 12, 2012. http://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/04.html (accessed 2015).

—. *Medical Subject Headings (MeSH®) in MEDLINE®/PubMed®: A Tutorial.* October 12, 2012. http://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/02.html (accessed 2015).

Pawson T, Gish GD, Nash P. "SH2 domains, interaction modules and cellular wiring." *Trends Cell Biol.*, 2001: 504-11.

Pawson T., Nash P. "Assembly of cell regulatory systems through protein interaction domains." *Science*, 2003.

Pawson T., Nash P. "Protein-protein interactions define specificity in signal transduction." *Genes Dev.*, 2000.

Pawson, T. *The Pawson Lab - Researching Signal Transduction.* 2006. http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=section&id=3&Itemid=64 (accessed 2015).

—. *The Pawson Lab - SH2 Domain.* 2005. http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id.

—. *The Pawson Lab - SH3 Domain.* 2005. http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=179 (accessed July 2015).

Pawson, T. *Domains - Listed.* February 24, 2006. http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=category&sectionid=3&id=45&Itemid=64 (accessed 2015).

Pawson, T. "Organization of cell-regulatory systems through modular-protein-interaction domains." *Philos Trans A Math Phys Eng Sci.*, 2003.

—. *The Pawson Lab - PDZ Domain.* 2005. http://pawsonlab.mshri.on.ca/index.php?option=com_content&task=view&Itemid=64&id=168.

Persaud A, Alberts P, Hayes M, Guettler S, Clarke I, Sicheri F, Dirks P, Ciruna B, Rotin D. "Nedd4-1 binds and ubiquitylates activated FGFR1 to control its endocytosis and function." (EMBO J.) 2011.

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B. "The tandem affinity purification (TAP) method: a general procedure of protein complex purification." *Methods*, 2001: 218-29.

Rao A., Bulusu G., Srinivasan R., Joseph T. "Protein-Protein Interactions and Disease." *InTech, DOI*, 2012.

Salah Z, Aqeilan RI. "WW domain interactions regulate the Hippo tumor suppressor pathway." *Cell Death Dis.*, 2011.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. "A new measure for functional similarity of gene products based on Gene Ontology." *BMC Bioinformatics*, 2006.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Res.*, 2003.

Sidhu, Sachdev. *Protein Binders.* 2014. http://sites.utoronto.ca/sidhulab/research2.html (accessed 2015).

Sinha, S. "On counting position weight matrix matches in a sequence, with application to discriminative motif finding." *Bioinformatics*, 2006.

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. "BioMart--biological queries made easy." *BMC Genomics*, Jan 2009.

Snel B, Lehmann G, Bork P, Huynen MA. "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." *Nucleic Acids Res.*, Sep 2008.

Sudol, M. "Structure and function of the WW domain." *Prog Biophys Mol Biol.*, 1996.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth ASantos A3, Tsafou KP3, Kuhn M4, Bork P5, Jensen LJ6, von Mering C7. "STRING v10: protein-protein interaction networks, integrated over the tree of life." *Nucleic Acids Res.*, Jan 2015.

Takeuchi K, Collier N. "Use of Support Vector Machines in Extended Named Entity Recognition." *COLING-02 proceedings of the 6th conference on Natural language learning*, Aug. 31, 2002: 1-7.

Thermo Fisher Scientific Inc. *Life Technologies.* 2015. https://www.lifetechnologies.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/co-immunoprecipitation-co-ip.html.

Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, Held HA, Appleton BA, Evangelista M, Wu Y, Xin X, Chan AC, Seshagiri S, Lasky LA, Sander C, Boone C, Bader GD, Sidhu SS. "A Specificity Map for the PDZ Domain Family." *PLoS Biol.*, 2008.

Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ. "iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence." *Database (Oxford)*, Oct 2010.

V. Srinivasa Rao, K. Srinivas, G. N. Sujini, G. N. Sunand Kumar. "Protein-Protein Interaction Detection: Methods and Analysis." *International Journal of Proteomics*, 2014.

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási AL, Vidal M. "An empirical framework for binary interactome mapping." *Nat Methods*, Jan 2009.

Vijaykumar YM, Vishal A. "Chromosomal Proximity of Genes as an Indicator of Functional Linkage." *SpringerBriefs in Systems Biology*, July 28, 2013: 33-42.

Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, Birge B, Cowburn D, Hanafusa H, Mayer BJ, Overduin M, Resh MD, Rios CB, Silverman L, Kuriyan J. "Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides." *Nature* 358 (August 1992): 646 - 653.

Walhout AJ, Vidal M. "High-throughput yeast two-hybrid assays for large-scale protein interaction mapping." *Methods*, 2001: 297-306.

Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. "A new method to measure the semantic similarity of GO terms." *Bioinformatics*, 2007: 1274-81.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function." *Nucleic Acids Res.* , 2010.

Wilks, Andrew F., and Ailsa G. Harpur. *Intracellular Signal Transduction.* 1996.

Wintjens R, Wieruszeski JM, Drobecq H, Rousselot-Pailley P, Buée L, Lippens G, Landrieu I. "1H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides." *J Biol Chem.* , 2001.

Wu C, Ma MH, Brown KR, Geisler M, Li L, Tzeng E, Jia CY, Jurisica I, Li SS. "Systematic identification of SH3 domain-mediated human protein-protein interactions by peptide array target screening. ." *Proteomics*, 2007.

You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP, Ji Z. "Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set." *BMC Bioinformatics*, 2014.

Yu FX, Guan KL. "The Hippo pathway: regulators and regulations." *Genes Dev.* , 2013.

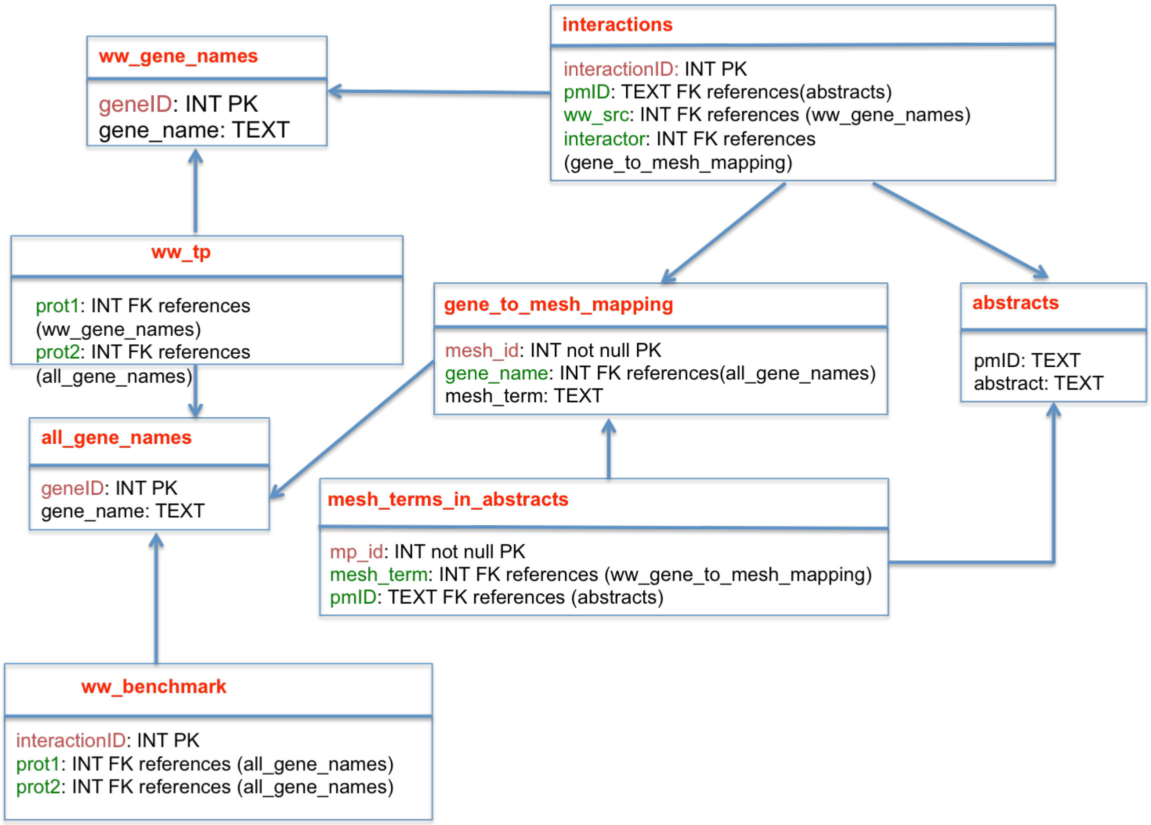
Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J, Liang Y, Yao L, Kasprzyk A. "BioMart: a data federation framework for large collaborative projects." *Database (Oxford)*, Sep 2011.

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. "Structure-based prediction of protein-protein interactions on a genome-wide scale." *Nature*, Oct 2012.

Zhang Y, Chang C, Gehling DJ, Hemmati-Brivanlou A, Derynck R. "Regulation of Smad degradation and activity by Smurf2, an E3 ubiquitin ligase." *Proc Natl Acad Sci U S A.*, Jan 2001.

Zhou MM, Ravichandran KS, Olejniczak EF, Petros AM, Meadows RP, Sattler M, Harlan JE, Wade WS, Burakoff SJ, Fesik SW. "Structure and ligand recognition of the phosphotyrosine binding domain of Shc." *Nature*, 1995.

Appendix A – Database ER diagram



ww_gene_names

geneID: INT PK
gene_name: TEXT

interactions

interactionID: INT PK
pmID: TEXT FK references(abstracts)
ww_src: INT FK references (ww_gene_names)
interactor: INT FK references (gene_to_mesh_mapping)

ww_tp

prot1: INT FK references (ww_gene_names)
prot2: INT FK references (all_gene_names)

gene_to_mesh_mapping

mesh_id: INT not null PK
gene_name: INT FK references(all_gene_names)
mesh_term: TEXT

abstracts

pmID: TEXT
abstract: TEXT

all_gene_names

geneID: INT PK
gene_name: TEXT

mesh_terms_in_abstracts

mp_id: INT not null PK
mesh_term: INT FK references (ww_gene_to_mesh_mapping)
pmID: TEXT FK references (abstracts)

ww_benchmark

interactionID: INT PK
prot1: INT FK references (all_gene_names)
prot2: INT FK references (all_gene_names)

Appendix B – Ensembl repository search results, using the Biomart query engine:

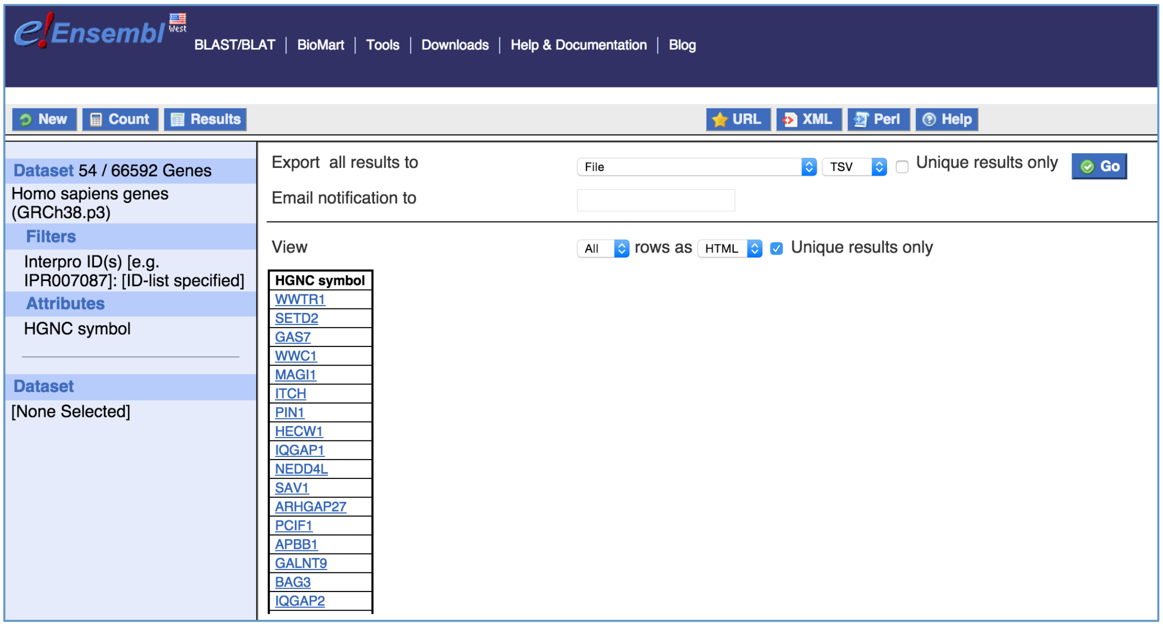
http://www.ensembl.org/biomart/martview/

Search filters:

Database: **Ensembl genes 81**

Dataset: **Homo Sapiens Genes (GRCh38.p3)**

Protein domain: **IPR001202** (corresponds to WW domain InterPRO ID)


e! Ensembl
West
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog
New
Count
Results
URL
XML
Perl
Help
Go

Dataset 54 / 66592 Genes
Homo sapiens genes (GRCh38.p3)

Filters
Interpro ID(s) [e.g. IPR007087]: [ID-list specified]

Attributes
HGNC symbol

Dataset
[None Selected]

Export all results to: File, TSV, □ Unique results only

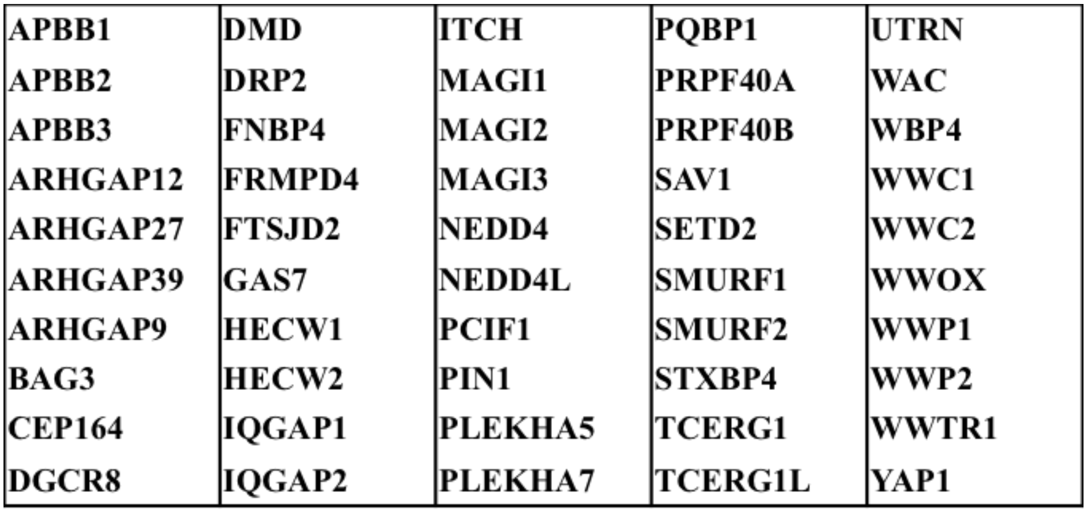
Email notification to

View: All rows as HTML ☑ Unique results only

HGNC symbol
WWTR1
SETD2
GAS7
WWC1
MAGI1
ITCH
PIN1
HECW1
IQGAP1
NEDD4L
SAV1
ARHGAP27
PCIF1
APBB1
GALNT9
BAG3
IQGAP2

Appendix C – Ensembl repository search results for WW domain

50 WW domain-containing proteins (`ensembl_ww_sorted.txt`)



APBB1
APBB2
APBB3
ARHGAP12
ARHGAP27
ARHGAP39
ARHGAP9
BAG3
CEP164
DGCR8
DMD
DRP2
FNBP4
FRMPD4
FTSJD2
GAS7
HECW1
HECW2
IQGAP1
IQGAP2
ITCH
MAGI1
MAGI2
MAGI3
NEDD4
NEDD4L
PCIF1
PIN1
PLEKHA5
PLEKHA7
PQBP1
PRPF40A
PRPF40B
SAV1
SETD2
SMURF1
SMURF2
STXBP4
TCERG1
TCERG1L
UTRN
WAC
WBP4
WWC1
WWC2
WWOX
WWP1
WWP2
WWTR1
YAP1

Appendix D – PubMed abstracts in XML format

File **pubmed_result.xml**

Sample:

```
<?xml version="1.0"?>
<data>
<PubmedArticle>
        <PMID Version="1">24008736</PMID>
            <ArticleTitle>WWOX suppresses autophagy for inducing
apoptosis in methotrexate-treated human squamous cell
carcinoma.</ArticleTitle>
                <AbstractText>Squamous cell carcinoma (SCC) cells
refractory to initial chemotherapy frequently develop disease
relapse and distant metastasis. We show here that tumor
suppressor WW domain-containing oxidoreductase (WWOX) (also named
FOR or WOX1) regulates the susceptibility of SCC to methotrexate
(MTX) in vitro and cure of SCC in MTX therapy. MTX increased WWOX
expression, accompanied by caspase activation and apoptosis, in
MTX-sensitive SCC cell lines and tumor biopsies. Suppression by a
dominant-negative or small interfering RNA targeting WWOX blocked
MTX-mediated cell death in sensitive SCC-15 cells that highly
expressed WWOX. In stark contrast, SCC-9 cells expressed minimum
amount of WWOX protein and resisted MTX-induced apoptosis.
Transiently overexpressed WWOX sensitized SCC-9 cells to
apoptosis by MTX. MTX significantly downregulated autophagy-
related Beclin-1, Atg12-Atg5 and LC3-II protein expression and
autophagosome formation in the sensitive SCC-15, whereas
autophagy remained robust in the resistant SCC-9.
Mechanistically, WWOX physically interacted with mammalian target
of rapamycin (mTOR), which potentiated MTX-increased
phosphorylation of mTOR and its downstream substrate p70 S6
kinase, along with dramatic downregulation of the aforementioned
proteins in autophagy, in SCC-15. When WWOX was knocked down in
SCC-15, MTX-induced mTOR signaling and autophagy inhibition were
blocked. Thus, WWOX renders SCC cells susceptible to MTX-induced
apoptosis by dampening autophagy, and the failure in inducing
WWOX expression leads to chemotherapeutic drug
resistance.</AbstractText>
</PubmedArticle>
</data>
```

Appendix E – A Python script to parse downloaded abstract search results in XML format and store in the database

```python
#!/usr/bin/python

import xml.etree.ElementTree as ET

import MySQLdb

# RUNS ON LARISA-DEV
db = MySQLdb.connect(host="localhost", user="root",
passwd="password", db="binding_site")
cursor = db.cursor()

tree = ET.parse('pubmed_result.xml')
root = tree.getroot()

for child in root:
    for c2 in child:
        if c2.tag == 'PMID':
            pmID = c2.text.strip()
        elif c2.tag == 'ArticleTitle':
            title = c2.text.strip()
        elif c2.tag == 'AbstractText':
            abstract = c2.text.strip()

cursor.execute("INSERT INTO abstracts(pmID, title,
abstract) VALUES('" + pmID + "', '" + title.replace("'",
"''") + "', '" + abstract.replace("'", "''") + "')")
```

Appendix F – iRefIndex PPI dataset, used as the benchmark set in PPI prediction script (ww_benchmark.tsv)

Sample:

alias a	alias b
crogid:EOGRPafDKt63MMyDOHF7BVg3Z6g|icrogid:1217575	uniprotkb:PIN1_HUMAN|entrezgene/locuslink:PIN1|crogid:FICkSpcBgvSDFW4iY1Etik85xNo9606|icrogid:1894250
crogid:Nge6q/DbfBB/c7EIeD0zSDci2NU|icrogid:947444	uniprotkb:SMUF2_HUMAN|entrezgene/locuslink:SMURF2|crogid:Ch9UvXcmtb2iD8i9DVKGCe77J7k9606|icrogid:1393325
crogid:WfcEVi/FMyvX8rgahjZYttRQ2Qs|icrogid:1099782	uniprotkb:DGCR8_HUMAN|entrezgene/locuslink:DGCR8|crogid:BpQaVWnWsgh5SDFrKplf7/9aWQw9606|icrogid:1271497
crogid:jUtJYlvMjurxZ7r7noeUQYVU4wQ|icrogid:886865	uniprotkb:ITCH_HUMAN|entrezgene/locuslink:ITCH|crogid:JyySOfI+ZHOI1mvdxWDGN6njgCY9606|icrogid:2639679
crogid:xALY6D9Likn8xKzeeWLJ/ChtkqI|icrogid:1114318	uniprotkb:DGCR8_HUMAN|entrezgene/locuslink:DGCR8|crogid:BpQaVWnWsgh5SDFrKplf7/9aWQw9606|icrogid:1271497
entrezgene/locuslink:ACCN3|crogid:SXlXB6ANoi3b+vd3LFnHMYFCqok9606|icrogid:4117748	entrezgene/locuslink:MAGI1|crogid:WN0iaCIXUObtJMGO7eRlQdaGDSk9606|icrogid:11793640
entrezgene/locuslink:AIMP1|crogid:UtCeSRjFyuuEHUxVF2HkG9RA55k9606|icrogid:4425202	uniprotkb:SMUF2_HUMAN|entrezgene/locuslink:SMURF2|crogid:Ch9UvXcmtb2iD8i9DVKGCe77J7k9606|icrogid:1393325
entrezgene/locuslink:APBB3|crogid:U1jEo7ZPG53cA/BcJs/BMGo/vWE9606|icrogid:4304478	entrezgene/locuslink:APLP1|crogid:6PKn8miZNOLOS3c4MjZFchs1yaQ9606|icrogid:715083
entrezgene/locuslink:APBB3|crogid:U1jEo7ZPG53cA/BcJs/BMGo/vWE9606|icrogid:4304478	uniprotkb:A4_HUMAN|entrezgene/locuslink:APP|crogid:HatR8w8+mNjtjr7s5+bAwaStpmk9606|icrogid:2186087
entrezgene/locuslink:APBB3|crogid:U1jEo7ZPG53cA/BcJs/BMGo/vWE9606|icrogid:4304478	uniprotkb:APLP2_HUMAN|entrezgene/locuslink:APLP2|crogid:WBmrQxD6+MjRPRNHt63XDIs9++o9606|icrogid:4664969
entrezgene/locuslink:ARHGAP27|crogid:E5nmKdVVnfW2ee8Uc9nb6Nb3rkA9606|icrogid:14007633	rogid:8UhNuJtumlKtKerMGo2BYQznJ9U9606|crogid:8UhNuJtumlKtKerMGo2BYQznJ9U9606|icrogid:9993265|-

Appendix G – A Python script to detect co-occurrence between WW domain-containing proteins and other human proteins in PubMed abstracts

```
#!/usr/bin/python

#################################################################
##################################
# RERUN W/O CV, PURE CO-OCCURRENCE
# ABSTRACTS ARE FROM 2012
# RUNS ON LARISA-DEV, DATABASE binding_site

# EXCEPTIONS: NES, INS, NTS, TES, TRO, EFS - Deleted them from
the human gene set and predictions table
#################################################################
##################################

import MySQLdb
import shlex, subprocess

# RUNS ON LARISA-DEV
db = MySQLdb.connect(host="localhost", user="root",
passwd="password", db="binding_site")
cursor = db.cursor()

# WW genes
cursor.execute("SELECT * FROM ww_genes")
results_ww = cursor.fetchall()

ww_genes = []

for r_ww in results_ww:
    ww_gene = r_ww[0]
    ww_genes.append(ww_gene)

wg_set = set(ww_genes)

# ALL genes
cursor.execute("SELECT * FROM genes")
results_g = cursor.fetchall()

genes = []

for r_g in results_g:
    gene = r_g[0]
    genes.append(gene)

cursor.execute("SELECT * FROM ww_abstracts_unique")
abst_res = cursor.fetchall()

abstrDict = {}

for a_res in abst_res:
    pm_ID = a_res[0]
    abstr = a_res[1]
    print pm_ID

    abstrDict[pm_ID] = abstr

s_genes = set(genes)

punctuation = [".", ",", ":", ";"]

#print abstrDict.keys()

for pmID in abstrDict.keys():

    if pmID != "":
        abstract = abstrDict[pmID]
        #print abstract

        for ww_gene in wg_set:
            #print 'WW gene: ' + ww_gene.lower()
            #print abstract.lower()

            combos = []

            # whole-word
            combos.append(" " + ww_gene.lower() + " ")

            # CANNOT use this, b/c gene 'INS' would return a
match for 'proteins' or 'domains' and gene 'NTS' matches
'variants' or 'elements', 'IDE' would match 'nucleotide'
            #combos2.append(g.lower() + " ")

            # and this would probably match 'insert'
            #combos2.append(" " + g.lower())

            # Dashes

            # And this matches REL to "-related"!!
            #combos2.append("-" + g.lower())

            combos.append("-" + ww_gene.lower() + " ")
            for p in punctuation:
                combos.append("-" + ww_gene.lower() + p)

            combos.append(" " + ww_gene.lower() + "-")
            combos.append("-" + ww_gene.lower() + "-")

            # Brackets
            combos.append("(" + ww_gene.lower() + ")")

            # Slashes
            combos.append("/" + ww_gene.lower())
            combos.append(ww_gene.lower() + "/")

            for p in punctuation:
                combos.append("/ " + ww_gene.lower() + p)

            for p in punctuation:
                combos.append(ww_gene.lower() + p)

            s_comb = set(combos)
            #print `s2_comb`

            for s in s_comb:
                if abstract.lower().find(s.lower()) >= 0:
                    print ww_gene + " found in abstr " + pmID

                    for g in s_genes:
                        if g != ww_gene and len(g) >= 3:

                            combos2 = []

                            # whole-word
                            combos2.append(" " + g.lower() + " ")

                            # CANNOT use this, b/c gene 'INS'
would return a match for 'proteins' or 'domains' and gene 'NTS'
matches 'variants' or 'elements', 'IDE' would match 'nucleotide'
                            #combos2.append(g.lower() + " ")

                            # and this would probably match
'insert'
                            #combos2.append(" " + g.lower())

                            # Dashes

                            # And this matches REL to "-
related"!!
                            #combos2.append("-" + g.lower())
                            combos2.append("-" + g.lower() + " ")

                            for p in punctuation:
                                combos2.append("-" + g.lower() +
p)

                            combos2.append(" " + g.lower() + "-")
                            combos2.append("-" + g.lower() + "-")

                            # Brackets
                            combos2.append("(" + g.lower() + ")")

                            # Slashes
                            combos2.append("/" + g.lower())
                            combos2.append(g.lower() + "/")

                            for p in punctuation:
                                combos2.append("/ " + g.lower() +
p)

                            for p in punctuation:
                                combos2.append(g.lower() + p)

                            s2_comb = set(combos2)

                            for s2 in s2_comb:
                                if
abstract.lower().find(s2.lower()) >= 0:
                                    print ww_gene + " and " + g +
" were found in abstract " + pmID

                                    # check reverse interaction!!
                                    if g in wg_set:
                                        cursor.execute("SELECT *
FROM co_occ_rerun_new_abstr WHERE gene1=" + `g` + " AND gene2=" +
`ww_gene` + " AND pmID=" + `pmID`)
                                        result =
cursor.fetchone()

                                        if not result:

cursor.execute("INSERT INTO co_occ_rerun_new_abstr(pmID, gene1,
gene2) VALUES(" + `pmID` + ", " + `ww_gene` + ", " + `g` + ")")
                                            break
                                    else:
                                        # still check if recorded
                                        cursor.execute("SELECT *
FROM co_occ_rerun_new_abstr WHERE gene2=" + `g` + " AND gene1=" +
`ww_gene` + " AND pmID=" + `pmID`)
                                        result =
cursor.fetchone()

                                        if not result:

cursor.execute("INSERT INTO co_occ_rerun_new_abstr(pmID, gene1, 
gene2) VALUES(" + `pmID` + ", " + `ww_gene` + ", " + `g` + ")")
                                        break
                    break
```

Appendix H – A Python script to compare predictions to benchmark and detect TPs and FPs

```
#!/usr/bin/python

import MySQLdb

# RUNS ON LARISA-DEV
db = MySQLdb.connect(host="localhost", user="root",
passwd="password", db="binding_site")
cursor = db.cursor()

outfile = open('co_occ_rerun_verification.tsv', 'w')

cursor.execute("SELECT * FROM ww_new_bm")
results = cursor.fetchall()

benchmark = []

for result in results:
    gene1 = result[0]
    gene2 = result[1]

    tup1 = (gene1, gene2)

    benchmark.append(tup1)

print `benchmark`


cursor.execute("SELECT * FROM co_occ_rerun_new_abstr")
results2 = cursor.fetchall()

for result2 in results2:

    pmID = result2[0].strip()

    gene1 = result2[1].strip()
    gene2 = result2[2].strip()

    b_tup = (gene1, gene2)
    b_tup_rev = (gene2, gene1)

    outfile.write(pmID + '\t' + gene1.strip() + '\t' +
gene2.strip() + '\t')

    if b_tup not in benchmark and b_tup_rev not in benchmark:
        outfile.write("FP\n")

    elif b_tup in benchmark or b_tup_rev in benchmark:
        outfile.write("TP\n")
```

Appendix I – Python script to retrieve MeSH terms for every human protein-coding gene.

```
   #!/usr/local/bin/python

   import re
   import string
   import EUtils
   import urllib
   import urllib2

   from Bio import Entrez

   from EUtils import HistoryClient

   Entrez.email = 'molhovsky@gmail.com'

   # Oct. 2, 2012: Search ENTIRE human genome, WW-prots AND
their interactors
   genes_file = open("approved_hgnc_symbol_biomart.txt", 'r')

   ww_prots = []

   outfile = open("mesh_results.txt", 'w')

   url = "http://www.ncbi.nlm.nih.gov/mesh/"

   ret_max = 1000

   num_abstr = 1

   pmids = []         # list of Pubmed IDs
   pubmedDict = {}    # dictionary: geneSymbol => [pmIDs]
   search_str = ""

   f_err1 = open("no_mesh_ids.txt", 'w')
   f_err2 = open("too_many_mesh_ids.txt", 'w')

   p_set = []

   for line in genes_file.readlines():
    prot = line.strip()
    p_set.append(prot)

   pSet = set(p_set)
   non_human_mesh = open('non_human_mesh.txt', 'w')
   gen_err_file = open('mesh_err.txt', 'w')

   for prot in pSet:
     try:
           handle = Entrez.esearch(db="mesh", term='"' + prot +
'"', rettype='text', retmax=ret_max)
           record = Entrez.read(handle)
           mesh_ids = record ["IdList"]

           if len(mesh_ids) == 0:
                f_err1.write("No Mesh IDs for " + prot + '\n')

           elif len(mesh_ids) <= 5:
                for uid in mesh_ids:
                     try:
                          #handle2 =
urllib.urlopen("http://www.ncbi.nlm.nih.gov/mesh?term=" + uid +
"[uid]")
                          handle2 = urllib.urlopen(url + uid)
                          record2 = handle2.read()

                          s1 = record2.find("[Supplementary
Concept]")

                          i = s1
                          c = record2[i]

                          while c != ">":
                               i -= 1
                               c = record2[i]

                          mesh_term = record2[i+1:s1]

                          if mesh_term.lower().strip() ==
prot.lower().strip() + " protein, human":
                               outfile.write(prot + '\t' +
mesh_term + '\n')

                          else:
                               non_human_mesh.write(prot + '\t'
+ mesh_term + '\n')

                     except urllib2.URLError:
                          gen_err_file.write(prot + '\n')
                          Entrez.email = 'gingerbraid@yahoo.com'
                          continue
           else:
                f_err2.write("Too many IDs: " + prot + '\n')
```

```
    except RuntimeError:
        gen_err_file.write(prot + '\n')
        Entrez.email = 'olhovsky@lunenfeld.ca'
        continue

    except IOError:
        gen_err_file.write(prot + '\n')
        Entrez.email = 'olhovsky@lunenfeld.ca'
        continue

    except urllib2.URLError:
        gen_err_file.write(prot + '\n')
        Entrez.email = 'olhovsky@lunenfeld.ca'
        continue
```

Appendix J – Python script for retrieving all MeSH abstracts in MEDLINE format and identifying the MeSH terms indexed in them as Supplementary Concepts.

```
#!/usr/local/bin/python

import re
import string
import EUtils

from Bio import Entrez

from EUtils import HistoryClient
from Bio import Entrez

Entrez.email = 'molhovsky@gmail.com'

ww_prots_file = open("ensembl_ww_sorted.txt", 'r')     # all WW-
containing proteins (51 total)
ww_prots = []

outfile = open("abstracts_EUtils.txt", 'w')

ret_max = 1000

num_abstr = 1

err_file = open('over_1000_abstracts.txt', 'w')

mesh_dict = {}  # prot, [MeSH Terms] query string

mesh_outfile = open("mesh_terms.txt", 'w')

for line in ww_prots_file.readlines():
    prot = line.strip()

    mesh_terms = "http://www.ncbi.nlm.nih.gov/pubmed?term="

    handle = Entrez.esearch(db="pubmed", term=prot,
retmax=ret_max)

    record = Entrez.read(handle)
    max_ret = int(record["Count"])

    if max_ret > ret_max:
     print "More than 1000 entries for " + prot

     err_file.write(prot + '\t' + `max_ret` + '\n')
     continue

    for pmID in record["IdList"]:

     h2 = Entrez.efetch(db="pubmed", id=pmID,
rettype="abstract", retmode="xml")

     for line in h2.readlines():
           ind1 = line.find("<AbstractText>")
           ind2 = line.find("</AbstractText>")

           if ind1 > 0 and ind2 > 0 and ind2 > ind1:
                outfile.write(`num_abstr` + ". " +
line[ind1+len("<AbstractText>"):ind2] + "\n\n")
                num_abstr += 1

           # Find MeSH terms:
           if line.strip().find('<DescriptorName') == 0:

                m_ind1 = line.strip().find(">")
                m_ind2 = line.strip().find('</DescriptorName>')

                if m_ind1 > 0 and m_ind2 > 0 and m_ind2 >
m_ind1:
                      mesh_descr = line.strip()[m_ind1+1:m_ind2]

                      if mesh_terms ==
"http://www.ncbi.nlm.nih.gov/pubmed?term=":
                           mesh_terms = '"' + mesh_descr +
'"[MeSH Terms]'
                      else:
                           mesh_terms += " AND " + '"' +
mesh_descr + '"[MeSH Terms]'

     mesh_dict[prot] = mesh_terms

     mesh_outfile.write('\n' + prot + '\n' + mesh_terms +
'\n\n')
```