Semantic modelling of protein-protein interactions, their prediction and evaluation

Laleh Kazemzadeh

A thesis submitted to
The College of Medicine, Nursing and Health Sciences (CMNHS)
Regenerative Medicine Institute (REMEDI)
National University of Ireland, Galway

In partial fulfilment of the requirement for the
degree of Doctor of Philosophy

Supervised By

Prof. Frank Barry
&
Prof. Dietrich Rebholz-Schuhmann

February 2018

Declaration of Authorship

I hereby declare that this thesis is entirely my own work and where I have consulted the published work of others, such work has been cited and acknowledged in the text. I have not obtained a degree in this university, or elsewhere on the basis of this work.

Signed: Laleh Kazemzadeh

Date: February 22, 2018

"Whether it be the sweeping eagle in his flight, or the open apple-blossom, the toiling work-horse, the blithe swan, the branching oak, the winding stream at its base, the drifting clouds, over all the coursing sun, form ever follows function, and this is the law. Where function does not change, form does not change."

Louis Sullivan

Abstract

The amount of biomedical data produced by DNA-sequencing, by curated knowledge on disease mechanisms and treatments, by results from biochemical and pharmaceutical research and by many other data generation studies is escalating at an unconstrained pace. However, this wealth of biomedical data is a precious resource for integrative research studies which draw conclusions through the analysis of all the heterogeneous data for knowledge discovery. One biomedical research domain is the prediction of protein-protein interactions relying on different sources of data which *a priori* may not directly expose data for protein interactions but may hold hidden information to identify novel interaction candidates. However, data integration and in addition data interpretation have to overcome a number of hurdles, which result from the characteristics of the biomedical data sources, including challenges from data diversity in protein namings, data consistency, analogy, availability and interoperability.

The aim of this thesis is harnessing the capability of *big biomedical data* by integrating its artifacts through the application of Semantic Web and the Linked Data principles for the final goal of predicting protein-protein interactions from the data. A semantic model for protein-protein interaction networks has been developed in this work which is used to identify explicit knowledge on protein interactions. This model is based on protein traits which have been extracted from publicly available biomedical data sources.

The research work in this thesis has led to the integration of descriptive features of proteins from public reference data sources denoting known interactions and subsequently initiated the prediction of novel interactions. The prediction model included novel attributes such as the genomic location of the genes and their immediate neighbouring network for each protein. Through the integration of these features, a Naïve Bayes approach achieved a prediction accuracy close to 94% measured against a gold standard of known protein-protein interactions.

The semantic integration of the biomedical data covers protein data and their interaction networks. This approach used state of the art integration techniques based on Linked Data principles relying on a basic ontology. The data is exposed as visual analytics platform (called LinkedPPI) optimised for intuitive data exploration.

A selection of predicted protein interactions has then been validated experimentally through laboratory experiments in order to demonstrate validation of the predicted interactions. The positive outcomes of the experimental validation demonstrate that the prediction model and the data integration form an effective means for the selection of most relevant but yet unknown protein interaction candidates.

Acknowledgements

I would like to thank my supervisors Prof.Frank Barry and Prof.Dietrich Rebholz-Schuhmann for their support, guidance and sharing their expertise with me throughout my PhD. I am grateful to Prof.Frank Barry for believing in me and giving me the opportunity to become a better scientist. His patience and encouragement helped me to reach to this point. I am thankful to Prof.Dietrich Rebholz-Schuhmann for his day to day support and passion which inspired me to push beyond the constrains. I am fortunate to have them both as my mentors.

I would like to thank all members of my graduate research committee Prof.Stefan Decker, Dr.Mary Murphy, Dr.Oya Beyan, Dr.Ratnesh Sahay for their constructive advices.

Special thanks to my colleague in Insight Centre for Data Analytics in Galway in particular Bahareh, Lukasz, Sanaz, Alessandra, Brian, Narumal, Achille, Sonya, Souleiman, Fadi, Gofran, Mr.Wassim, Soheila, Arindam and Myriam for creating the most amazing memories of my Galway life.

Thanks to all my fellow postgraduate students in Stem Cells and Orthobiology group for cheering me up when lab work got tough. Thank you Cathy, Claire, Sean, Swarna, Maojia, Yvonne and Patrizio for making the transition form simulation scientist (that is how Frank thinks about me) to a real scientist so smoothly. Georgina you are the best of all.

A special word of thank to those who I worked closely with over the course of my PhD, Maulik, Rezaul and Julien, for sharing their technical knowledge with me.

Last but certainly not the least, I would like to thank my family for supporting and encouraging me at all time and enduring my long absence to become a better version of me.

My research is partially supported by the Structured PhD in Simulation Science which is funded by the Programme for Research in Third Level Institutions (PRTLI) Cycle 5 and co-funded by the European Regional Development Fund. This thesis has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

Contents

Declaration of Authorship ii

Abstract v

Acknowledgements vii

List of Figures xii

List of Tables xiii

Abbreviations xiv

1 Introduction 1
- 1.1 General background . . . 1
  - 1.1.1 Biological networks . . . 2
  - 1.1.2 Importance and Application of Protein-Protein Interaction Networks 3
- 1.2 Biomedical data sources and *in silico* models . . . 5
  - 1.2.1 Biomedical data sources . . . 6
  - 1.2.2 Prediction of PPIs . . . 7
- 1.3 Limitations and challenges . . . 9
- 1.4 Experimental background . . . 10
  - 1.4.1 Mesenchymal Stem Cells (MSCs) . . . 11
  - 1.4.2 Cartilage biology and disorders . . . 12
  - 1.4.3 Cartilage components . . . 13
    - 1.4.3.1 CD44 . . . 13
    - 1.4.3.2 TNF-stimulated gene 6 protein (TSG-6) . . . 16
    - 1.4.3.3 Hyaluronan and proteoglycan link protein 1 (HAPLN1) . 18
- 1.5 Aims and objectives . . . 19

2 Related Works 21
- 2.1 Modelling of PPI predictions . . . 22
  - 2.1.1 Descriptive features . . . 22
    - 2.1.1.1 Sequence-based features . . . 22
    - 2.1.1.2 Concept based features . . . 23
  - 2.1.2 Prediction algorithms . . . 24

2.1.3 Evaluation and comparison of performances for state of the art approaches . . . 27
2.2 Standardisation of biomedical data . . . 30
2.3 Conclusion . . . 32

3 Prediction of Protein-Protein Interactions using 3D chromosomal locations 33
3.1 Introduction . . . 33
3.2 Methodology . . . 34
3.2.1 Prediction algorithm . . . 34
3.2.2 Gold standard dataset . . . 36
3.2.3 Feature selection . . . 38
3.2.3.1 Domain-domain interactions . . . 38
3.2.3.2 GO similarity . . . 39
3.2.3.3 Neighbouring network . . . 39
3.2.3.4 Genomic location . . . 40
3.2.4 Gene ontology enrichment analysis . . . 41
3.3 Results . . . 41
3.4 Discussion . . . 43

4 Semantic Integration 47
4.1 Introduction . . . 47
4.2 Methods . . . 49
4.2.1 Selection of relevant data sources . . . 49
4.2.1.1 Validated interactions as a dataset . . . 49
4.2.1.2 Protein complexes . . . 50
4.2.1.3 Gene expression . . . 51
4.2.1.4 Genomic locations . . . 51
4.2.1.5 Protein domains . . . 51
4.2.1.6 Gene co-occurrence . . . 52
4.2.2 Entity mappings . . . 52
4.2.3 The LinkedPPI domain specific model . . . 53
4.2.3.1 Description of concepts . . . 54
4.2.3.2 Data transformation . . . 55
4.3 Results . . . 55
4.3.1 Statistics on RDFized data . . . 55
4.3.2 Visualization in LinkedPPI . . . 56
4.3.3 Use Cases . . . 59
4.3.3.1 Use Case 1: PPI candidates based on domain-domain interactions . . . 59
4.3.3.2 Use Case 2: Domain-domain interactions candidates . . . 61
4.3.3.3 Use Case 3: Selective interactions between segments of the Human genome. . . . 61
4.3.4 Domain-domain interaction dataset . . . 63
4.4 Discussion . . . 64

5 Experimental Validation 66
5.1 Overview of the interaction network . . . 68

5.2 Methods and Materials 70
5.2.1 ATDC5 cell culture and expansion 70
5.2.1.1 Chondrogenic differentiation 70
5.2.1.2 Alcian Blue staining of ATDC5 70
5.2.2 Detection of Proteins Expression 71
5.2.2.1 Cell Harvest 71
5.2.2.2 Quantification of Proteins Using BCA Assay 71
5.2.2.3 List of Primary and Secondary Antibodies 71
5.2.2.4 Western blot (WB) 72
5.2.3 Co-Immunoprecipitation (Co-IP) 72
5.2.3.1 Protein visualisation by Enhanced Chemiluminescence . . 73
5.2.3.2 Immunostaining of mono layer ATDC5 74
5.2.3.3 Immunostaining of knee sections 74
5.3 Results 75
5.3.1 Morphology of ATDC5 cells in culture and chondrogenic differentiation 75
5.3.2 Quantification of protein concentration in ATDC5 76
5.3.3 Detection of protein expression in chondrogenesis 77
5.3.4 Detection of proteins co-localisation by immunofluorescence (IF) . 77
5.3.5 Detection of protein binding 79
5.4 Discussion 80

6 Conclusion **83**
6.1 Thesis summary 83
6.2 Contributions 84
6.3 Future research direction 85

Bibliography **87**

List of Figures

1.1 The average number of PPIs released by the BioGRID. 8
1.2 A schematic representation of healthy joint vs. OA. 13
1.3 A schematic representation of CD44. 14

3.1 Landscape of the model. 35
3.2 GO enrichment of chromosomes for biological process (P-value ≤ 0.05). . 42

4.1 LinkedPPI Architecture 50
4.2 Class Diagram of LinkedPPI Domain-specific Model 53
4.3 Searching HES1 protein using the PPI visualisation dashboard 56
4.4 Subgraph of HES1 PPI network based on GO terms 57
4.5 Illustration of the three use cases. 60

5.1 LinkedPPI output for CD44 search. 67
5.2 CD44 partners retrvied form LinkedPPI. 68
5.3 Network of interactions for CD44 and TSG-6 predicted by String. 69
5.4 Network of interactions between CD44, TSG-6, HAPLN1. 69
5.5 Monolayer cultured ATDC5 cells, 4x magnification. 75
5.6 Confirmation of chondrogenesis by Alcien blue staining for GAG on day 21. 75
5.7 Standard curve generated to determine protein concentrations of ATDC5 cells at Day 0 to Day 21 in both expansion and chondrogenic media. . . . 76
5.8 Expression of CD44, TSG-6, HAPLN1 proteins during expansion and chondogenesis. 77
5.9 Safranin O stained control knee section. 78
5.10 IF of ATDC5 culture monolayer on day 21, 60x magnification. 78
5.11 IF of collogenase treated mice knee section, 40x magnification, scale bar 50mM. 79
5.12 Co-IP of CD44 anti TSG-6 and anti HAPLN1 on day 0. 80
5.13 Co-IP of CD44 anti TSG-6 on day 21. 80
5.14 Co-IP of CD44 anti HAPLN1 on day 21. 81

List of Tables

1.1 List of data repositories which host PPI data. 7

2.1 Summary of performance of state of the art models. 29

3.1 Features and data sources. 38
3.2 Feature composition of different scenarios. 40
3.3 Average model performance on different feature scenarios. 44

4.1 Integrated data sources. 50

5.1 Western blot buffers. 73
5.2 Protein concentration. 76

Abbreviations

3did	The database of Three-Dimensional Interacting Domains
ACC	Accuracy
AIA	Antigen-Induced Arthritis
ATDC5	Murine Chondrogenic Cell Line Derived from Teratocarcinoma AT805
BCA	Bicinchoninic Acid Assay
BioGRID	Biological General Repository for Interaction Datasets
BP	Biological Process
C4S	Chondroitin-4-Sulphate
CC	Cellular Compartment
CFG	Consortium for Functional Glycomics
Co-IP	Co-Immunoprecipitation
CORUM	The Comprehensive Resource of Mammalian Protein Complexes
Crtl1	Cartilage-Linking Protein 1
dH2O	Deionized Water
D-PBS	Dulbeccos Phosphate Buffered Saline
DMEM	Dulbeccos Modified Eagle Medium
ECACC	European Collection of Cell Cultures
ECM	Extra Cellular Matrix
EDTA	Ethylenediaminetetraacetic Acid
EGF	Epidermal Growth Factor
FBS	Fetal Bovine Serum
FGF	Fibroblast Growth Factor
FN	False Negative
FP	False Positive

GAG	Glycosaminoglycans
GAP	Glycome Analytic Platform
GO	Gene Ontoloy
HA	Hyaluronic Acid
HAPLN1	Hyaluronan and Proteoglycan Link Protein 1
HGNC	HUGO Gene Nomencla- ture Committee
IαI	Inhibitor inter- α-Inhibitor
IF	Immunofluorescence
Ig-like	Immunoglobulin-like
IL-1	Interleukin-1
ITS	Insulin-Transferrin-Selenium
***k*NN**	k Nearest Neighbour
Mef2c	Myocyte Enhancer Factor 2C
Met	Mesenchymal-Epithelial Transition
MF	Molecular Function
MMP	Matrix Metalloproteinase
MSC	Mesenchymal Stem Cells
NB	Naïve Bayes
OA	Osteoarthritis
PBS	Phosphate-Buffered Saline
PCC	Pearson Correlation Coefficient
PDGF	Platelet-Derived Growth Factor
PMSF	Phenylmethanesulfonyl Fluoride
PPI	Protein-Protein Interaction
PSI-MI	The Proteomic Standards Initiative-Molecular Interaction
PSSM	Position-Specific Scoring Matrix
PTM	Post Translational Modification
PVDF	Polyvinylidene fluoride
RDF	Resource Description Framework
ReVealD	Real-time Visual Explorer and Aggregator of Linked Data
RF	Random Forest
RT	Room Temperature
Sens	Sensitivity

SOX9	Sex-Determining Region Y Box 9 Protein
Spec	Specificity
SPPS	Sequence-based Protein Partners Search
SVM	Support Vector Machine
TAN	Tree-augmented Naïve Bayes
TBS	Tris-Buffered Saline
TGF-β	Transforming Growth Factor Beta
TGP	Toxicogenomics Project
TN	True Negative
TNF	Tumor Necrosis Factor
TNFAIP	Tumor Necrosis Factor-Stimulated Gene-6 Protein
TP	True Positive
TR1,2	Tandem Repeat 1,2
TSG6	TNF-stimulated gene 6 protein
VCF	The Variant Call Format
WB	Western Blot
Y2H	Yeast 2 Hybrid

To those who made it happen...

Chapter 1

Introduction

Genomic data is generated at a higher pace than it is analysed. However, this data has a short life cycle since it is the answer to only one particular research question. Attaining higher analytic momentum as well as re-using data to address the diverse range of research questions has inspired engagement of more computational models. To a very high extent computational models are able to mimic the cellular behaviour given the complete and accurate data. However, despite all the advances in *-omics* technologies data is far from complete. On the other hand the wealth of available data provides the opportunity for the models to learn from the data itself in order to fill the gaps in our knowledge about cardinal cellular behaviours. Cellular behaviours are cell responses to the stimuli which are propagated via regulatory networks to and from protein-protein interaction (PPI) networks. Therefore, a thorough understanding of the dynamics of PPI networks is crucial in understanding these behaviours.

The focus of this thesis is the better discernment of PPI networks by recognising the novel protein traits. These traits were observed and extracted from existing data sources. In this thesis, I propose a framework for rehashing associated data in order to expand and facilitate more intuitive data exploration for the domain experts without extensive computational expertise. The current chapter provides the general biological background, genomic data sources and brief overview of computational models in the prediction of new PPIs.

1.1 General background

Proteins are the most vital molecules in a living system. They serve a diverse range of functions in essential biological processes in order to secure the stability and survival of

the cell. Proteins are built of 20 amino acids monomers, forming a linear chain. The long polymer forms the primary structure of the protein. Arrangements of amino acids along the polymer and their physio-chemical properties force the linear sequence to fold into a unique three dimensional structure. The primary sequence of a protein dictates its folding however this is not a one-to-one relation. Proteins with similar 3D conformation may have different primary sequences. From linear sequence to a three dimensional protein structure, proteins gain a vast array of functionality. Functions which are carried out by functional groups including alcohols, hydroxyl, methyl, carbonyl and carboxyl if brought together expand the range of proteins' activities.
In a living system proteins interact with each other either forming protein complexes or transmitting signals through the cascade of other proteins in response to internal and external clues. Thus, understanding the interplay between proteins will shed light on a better understanding of biological networks.

1.1.1 Biological networks

A biological network is a tightly controlled sequence of interactions between molecules in a cell which results to cellular changes in response in the wide spectrum of stimuli. A cell is continuously receiving intrinsic and extrinsic signals triggered by injuries, infections and stress which consequently prompt and activate proteins in order to pass on these signals. As an example, if the *stop* signal during cell division does not transmit properly it may turn a healthy cell into a cancerous one. The same takes place, if there is a disturbance in delivering the *programed death* command. Thus, it is crucial for the cell to maintain the balance and the sequence of the interactions between various elements of the cell regulation in order to ensure the activation of relevant pathways at any point in time.

Biological networks are far more complex than what was thought previously. The beginning and end of these networks are entangled to one another and results to their interplay in order to complete the designated tasks. Current technologies enable us to take snapshots of inter-connectivity of these networks in a static manner while the interactions between and within networks change dynamically over time and may differ from one cell type to another. Therefore, unveiling these connections fills the gaps and brings together the different pieces of the puzzle which then form the *bigger picture* of these networks and subsequently of the cellular processes.

The biological network denotes the interactions between molecules, DNA, proteins and genes. In particular, the interactions between proteins and DNA segments resulting into

gene expression form the *regulatory networks* and the physical interactions between pairs or multiple proteins form the *protein-protein interaction networks*.

The human genome is composed of around 24,000 genes and the protein products of the genes control the functionality of the cells. A regulatory network controls the temporal and spatial abundance of these products. Such a network is a collection of regulatory elements such as transcription factors which specifically bind to a region of DNA and lead to the expression of the genes residing in that region. A number of diseases result from the perturbations at the level of gene expression which then leads to the interruption in the protein interaction networks. Hence, unravelling the complexity of the regulatory and interaction networks will shed light on the mechanisms behind the initiation and progression of several diseases.

Proteins rarely act as single units but are always involved in a functional complex, i.e. in a network of interacting partners that forms a realistic and accurate model of biological processes in contrast to any pairwise representation of the interactions. Even though our knowledge of the complete repertoire of the interactome is far from complete, PPI networks still provide invaluable information for innovative research in the functional annotation of proteins, drug-target studies and pathway analysis.

1.1.2 Importance and Application of Protein-Protein Interaction Networks

The characterisation of unknown proteins is one essential and ambitious goal of computational biology mainly due to the vital role of proteins. The advances in genome sequencing technologies and in the identification of new proteins are by far faster than manual processing and experimentally annotating newly discovered proteins which broadens the gap between our knowledge of existing proteins and their crucial functions in the cell. One solution to such a challenge is the application of PPI and annotation of the unknown proteins in the context of their interacting partners. It has been shown that protein pairs share at least one molecular function and the number of times a function appears in the protein neighbourhood in the network can be used as an estimation for functional annotation [1]. Some studies differentiate between the immediate next neighbouring proteins and indirect or N hop neighbours away [2] in which a weight has been assigned to the annotation estimation. The common neighbour-based model is another approach which was proposed by Lin et al.[3] and is based on the assumption that proteins with common functionality would have common neighbours as well. This implies that the higher the number of shared partners is between two proteins, the higher is the chance that they have similar functionality. Zhang et al.[4] proposed a domain-context

approach. Based on their method, similar domain composition among neighbouring proteins is an indication of evolutionary closeness and comparable functionality between pairs. Based on aforementioned references PPI networks complementary to computational algorithms demonstrate a valuable resource in the characterisation of new proteins and their functionality.

A sequential chain of events in PPI networks is underlying biological phenotypes (e.g: spectrum of human diseases). Therefore, it is crucial to maintain this sequence as it is naturally programmed in the cell in order to indemnify cell viability and a healthy organism. Most of diseases are the result of disruption or perturbation in signalling networks. Even though genetic mutations are reported to be closely related to most cancer types these alterations in a gene sequence affect protein translation and lead to disruption of interactions in the signalling networks. From the network perspective, cancers are caused by more interacting with oncogenes and less with tumour suppressor genes.

In a similar fashion, drugs are designed to target one protein. More precisely they are designed to alter, inhibit or promote one of the protein's interactions [5]. Given the known few altered connections in the network, it is expected that drugs are designed in a way that they target only one specific edge of the network while the reality is different. Most of the drugs bind to more than one target protein in the network thus minimization of non-specific binding is required in order to optimise the drug's effectiveness. Hence, drug-target design highly benefits from a better understanding of PPIs and their interplay. On the other hand assignment of direct genotype-phenotype relations in most of the multi-factorial diseases (e.g: cancer, diabetes) is challenging. Thus, it is important to study the gene-disease causality relation as part of the PPI networks and not individually.

Assuming that modifications in PPI networks dictate the phenotypic variation, such changes are aggravated both by the changes in the level of abundance of the proteins and disconnection in its network of interactions.

In some diseases, multiple proteins are identified as drug targets therefore, multi-target drug designs are commonly used. Moreover, in diseases such as cancer where the drug's target adapts to the drug, minimise the effect of the drug and leading to disease recurrence, administration of combinatorial drugs diminishes the risk of drug resistance associated with the single drug as well as preventing the risk of drug-drug interactions. The underlying premise in network-based drug design approaches is the location and importance of the target protein in the network. It presumes that the more connected and centrally located the protein is the more essential is the protein, therefore it makes

a better candidate as a target. Accordingly targeting the hub proteins will have a destructive effect in cancer treatment as opposed to targeting the hub's neighbour in order to adjust slight perturbations in the network. Similarly, proteins at the cross-talk of pathways are better target candidates.

Aforementioned studies provide several lines of evidence on the importance of a better understanding of PPI's modulation and emphasis on their seminal therapeutic role in human diseases as well as functional annotation studies.

1.2 Biomedical data sources and *in silico* models

A wide range of experimental methods are able to detect the interaction between proteins and their partners. These methods can be classified as high-throughput or large-scale and low-throughput or small-scale methods. Generally, high-throughput methods detect a number of interactions as opposed to small-scale methods which detect the binary interaction between a given protein pair. Both high and low throughput methods have their advantages and disadvantages. The main drawback of high-throughput methods is the generation of a large number of false positive interactions thus it is less accurate. On the other hand, low-throughput methods are selective therefore they are more accurate. However, this accuracy comes with an increase in the laboratory time and labour cost.

The two most widely used gold standard techniques in detection of PPIs are co- immunoprecipitation and the yeast two-hybrid (Y2H) methodology. Co-immunoprecipitation is the most reliable small-scale method that is able to detect interactions between two or more proteins. In principal all the partners of the tagged protein are detectable by this method. However, this method is not able to detect the one-to-one interaction when the proteins form a complex which is its main drawback.

The yeast two-hybrid which was first introduced in 1989 [6] detects the interaction between a protein and its partners in the nucleus of yeast cells. The main disadvantages of this method is the generation of high number of false positive and possibly false negative interactions. Multiple reasons lead to such high false discovery rate including: a) the proteins under investigation might not be localised to yeast nucleus. In particular identification of the interactions between membrane and cytoplasmic proteins are more difficult [7]. b) The two proteins might not be expressed in the same cell type which increases the false positive rate. Taking into the account the limitation of the yeast two-hybrid method it is essential to verify the findings of this method with an additional methods such as affinity purification followed by mass spectrometry (AP/MS).

While Y2H identifies the binary interactions between a pair the AP/MS detects elements which are involved in a complex of proteins. The AP/MS method is able to identify protein complexes from cell lysate which provides a dynamic understanding of the proteins complexes in a native physiological environment, *in vivo*, as opposed to yeast nucleus [8]. However, AP/MS is also associated with high rate of false positive.

Large or small scale protein interaction methods generated large number of interactions which form the base of network analysis in the human interactome or other model organisms. By representing an interaction network in form of a graph it is possible to investigate the dynamics of these networks based on topological attributes of interactions networks including formation of modules. The graph topology including node degree [9], node distances [10, 11], structure [12] and modularity [13] represent the biological features of these networks. However, modularity has proved to be the most relevant attribute which can capture the dynamics of the protein interaction networks. CClusters of proteins encode multiple biological features: the centrality of the proteins, formation of complexes, and transient and permanent interactions. In an extensive study Han et al. [14] identified party and date hubs. The former are highly co-expressed with their partners indicating formation of protein complexes and the latter show varying expression profiles suggesting their transient role.

The interactions identified in experiment using either methods and used in any graphical network representation is available as raw data which requires processing, analysing and warehousing. Public data warehouses serve a wide range of researchers in the community to benefit from past research experiments. The next section is dedicated to introducing publicly available databases which host PPI data and *in silico* models of PPI predictions.

1.2.1 Biomedical data sources

Regardless of the method of identification, PPIs are generally hosted in publicly available repositories. Examples of such data sources are BioGRID [15], DIP [16], HPRD [17], MINT [18] and IntAct [19] which contain experimentally detected PPIs in form of binary or one-to-one records with literature references. On the other hand, KEGG [20], Reactome [21] and Pathway Commons [1]host information on pathways including their active elements, structures, overlaps and interplay. In addition, UniProt [22], Ensembl [23], Entrez-Gene [24] and Gene Ontology [25] offer sequence information, genomic localisation and cellular functionality of individual genes and proteins.

Table 1.1 shows a list of available biomedical data sources related to PPIs including their supporting data and data format. Evidently, the two most apparent data format in

[1] `http://www.pathwaycommons.org/`

these data sources are Tab-Separated Values (TSV) and Proteomics Standards Initiative-Molecular Interaction (PSI-MI). Nevertheless, neither of these standards and widely used data formats allow the interconnectedness of the data and hence create segregated and sporadic data.

TABLE 1.1: List of data repositories which host PPI data.

Databases	Data format	RDF distribution	Supporting data
BioGRID	TAB, PSI-MI	x	Literature
MINT	MITAB	x	Literature
DIP	PSI-MI, MITAB, XIN	x	IMEx consortium
IntAct	CSV, PSI-MI	x	Literature
MIPS	PSI-MI	x	Literature
I2D	PSI-MI	x	Literature
APID	TAB, PSI-MI	x	High throughput
Pathway Commons	RDF	✓	Multiple data sources
Reactome	BioPAX, PSI-MI, SBML, SBNG	✓	Multiple data sources

In addition to data sparsity, data duplication is observed in most of the data sources since most of the PPI data are extracted from the literature. *"Starting September 2013, MINT uses the IntAct database infrastructure to limit the duplication of efforts and to optimise future software development."* is stated on the home page of the MINT website. This statement demonstrates that data providers recognise the hindrances originating from the lack of connectivity. However, data repository convergence does not resolve the issues arising from augmenting PPI data and PPI related data. Data repositories like BioGRID apply the same principle by collecting the PPI data from various sources in order to prevent data duplicity. Nonetheless, the incorporation of the relevant data is still restrained.

Nevertheless, Reactome and Pathway Commons adopted the latest data format called Resource Description Framework (RDF). RDF supports interoperability among the data and is associated with the Linked Data concepts and Semantic Web technology, therefore it attracts the attention of researchers with integrative approaches interests.

1.2.2 Prediction of PPIs

Understanding the dynamics of PPI networks is a cardinal step in studying human diseases and drug developments at the molecular level. Hence, a thorough understanding of the signalling networks is essential. Owing to the advances in *-omics* technology vast amounts of PPI data have been accumulated. However, the main drawbacks of these techniques are the high false positive rate of interactions [26–28] and the generation of incomplete PPI networks. On the other hand, more accurate methods (e.g. Fluorescence resonance energy transfer (FRET)) which can detect direct interactions between pairs of

proteins are costly and time consuming, therefore the number of identified PPIs covers only a small fraction of the total number of existing PPIs in the higher organisms' interactome [29, 30]. The human genome contains more than 20,000 protein coding genes which brings the total number of binary interactions between pairs of proteins to 20,000!. Evidently, the experimental validation or rejection of these binary interactions is an unattainable task. However, proteins bind together selectively and leave signatures of their preferable partners behind. These footprints can be used in order to identify interaction patterns and deduce novel PPIs. These patterns might be between partners of a protein or among proteins with similar functional structure.

Figure. 1.1 shows the average growth of the number of identified PPIs provided by the BioGRID database over the last 10 years. There is an apparent strong increase in the accumulation of PPI data in this time interval which is an indicator of the increase in the amount of research concerned with proteins.

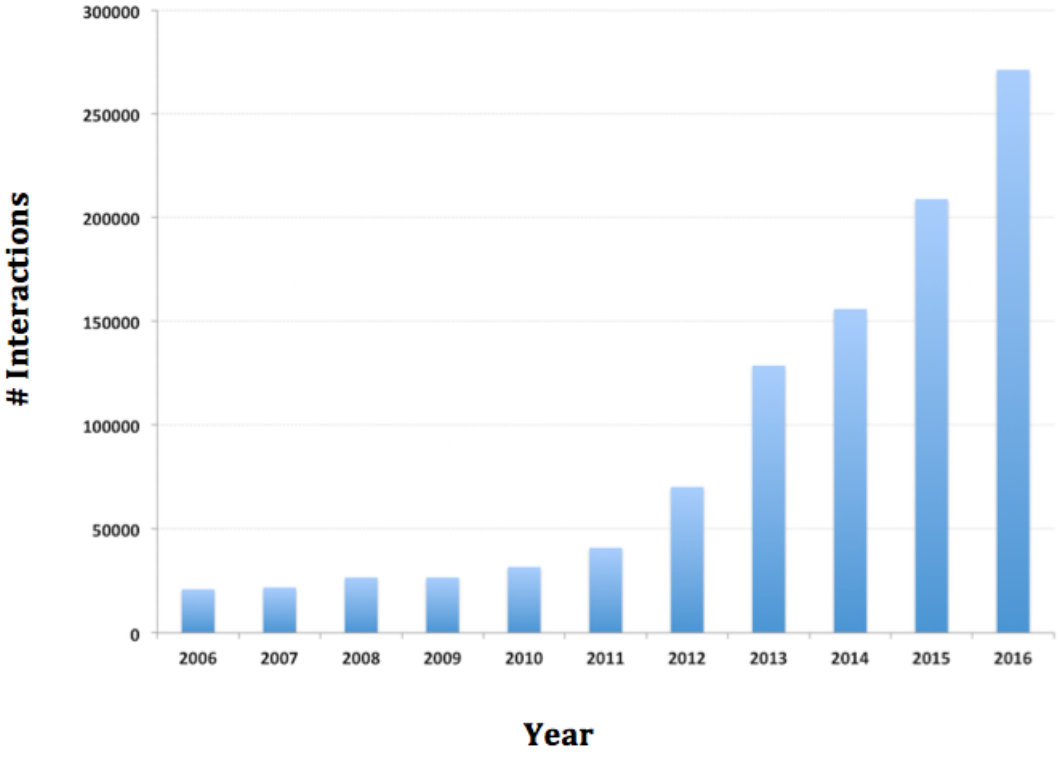


FIGURE 1.1: The average number of PPIs released by the BioGRID.

The abundant number of PPIs shown in figure. 1.1 is the main source in the identification of preferential interaction patterns among proteins. Hence, computational methods are required in order to recognise and highlight the underlying patterns which then will be used to predict potentially novel interactions.

Computational methods have been proposed for the prediction of PPIs which employ sequence information, domain composition, gene expression, ortholog interactions, phylogenetic conservation either as a single data source or a composition of multiple data sources in order to exploit the implicit information. Depending on the type of integrative data, the reliability of the integrated data sources and algorithms which have been applied, these computational models may return different levels of accuracy in PPI predictions. A more detail description and comparison of these models is described in Chapter 2.

1.3 Limitations and challenges

The enormous amount of biological data which has been generated as well as the tedious and costly task of experimental investigation highlights the need for computational models to assist the researchers in forming their hypothesis and drawing conclusions. In the analysis of PPI networks, PPIs themselves are the main source of data and form the backbone of the study. However, the integration of pertinent data such as gene expression, protein localisation, domain interactions has proved to be a valuable contribution towards the identification of the underlying patterns among the PPIs. These types of data are generally deposited in public data warehouses. Even though most of these data repositories represent data related to the PPIs or extended data relevant to the genes and proteins involved in these interactions, still data models and schema are highly diverse. This diversity stems from the lack of interoperability which by itself is due to the absence of criteria in data representation practices. Nonetheless, this data is generated in order to address particular research questions; therefore the type of laboratory methodologies, biological outcomes and consequently the data are diverse and appear to be irrelevant to the investigation of core PPI networks. Furthermore, the data representation and data schema is specific to their host repositories and may differ from one repository to another.

The main challenge is how to transfer the deluge of biomedical data to the knowledge that serves domain experts instead of being buried in the data warehouses. Biomedical data is shown to be generally isolated, sparse, repetitive and research question oriented. For example, multiple data repositories might refer to a specific protein, however, each of these repositories might use a specific identifier to this protein. Similarly, a specific gene may have multiple isoforms and these isoforms may have different identifiers in different repositories. In both cases, the link between two identifiers is missing, therefore leading to a gap in knowledge representation and consequently creating gaps in the integration of multiple data sources. In the above examples, both identifiers refer to either

proteins or genes while in an integrative study (e.g. the PPI prediction) heterogeneity of multi-dimensional data widens the gaps and thus hinders the ultimate capability of the integrative models.

Challenges in the integration of massive amount of heterogeneous data lead researchers to adopt a new generation of integrative methods based on Linked Data principles. Towards this end, several attempts have been made in the standardisation of the data through controlled vocabularies and guidelines but various hurdles still need to be overcome. The PSI-MI [31] is widely accepted by the community for the modelling of biological networks. Even though some of the data sources are represented using the PSI-MI format, there is no decipherable interconnectedness between them which highlights the potential of Linked Data technologies in this domain. Linked Data constructs the *bridge* between these data sources by making use of *ontologies* to achieve interoperability between data sources.

Ontologies are a set of terminologies and vocabularies that describe the raw data and relations among the data. Ontologies have been used in data integration, data exchange and reasoning. They vary in terms of size, coverage of the domain terminologies and their adoption in the research community. On the other hand, ontologies contain several non-domain specific terms which introduce unnecessary complexity in addressing research problems such as the analysis of PPI networks. Therefore, the application of a *domain specific vocabulary* is a more suitable approach for such focused research questions. Domain specific vocabularies are limited to the domain terminologies, small in size and designed to address specific questions while maintaining the interoperability of the data model.

Over and above all, ontologies or domain specific vocabularies not only resolve the heterogeneity of the data but also provide a solid ground for exchange, sharing and re-use of experimental data among the research community. Despite all the current efforts in data standardisation and ontology development, the lack of interoperability among biomedical data is still apparent. This underlines the need for wider adoption of openness, unification and sharing of the data amongst domain researchers. This might be more achievable by breaking down the more general and extensive ontologies into limited but more focused domain related vocabularies.

1.4 Experimental background

The current section presents an introduction to the experimental aspects of this thesis. The evaluation of the proposed model demonstrates the interaction between multiple

proteins in mesenchymal stem cells (MSCs) isolated from bone marrow of mice. These proteins are components of extracellular matrix (ECM) and are involved in maintaining the cartilage structures. Thus it is essential to describe these concepts beforehand in order to attain a better understanding of the experimental outcome in the following chapters. It is noteworthy that the choice of MSCs and selection of proteins were solely based on the expertise and interests of the laboratory. However, in theory the proposed model could be applied to any protein and its potential partners.

1.4.1 Mesenchymal Stem Cells (MSCs)

Mesenchymal stem cells (MSCs) are a subtype of stromal cells with self-renewal and multilineage differentiation potentials. These cells are found across a wide range of tissues. MSCs were first isolated from bone marrow by Friedenstein et. al [32] and later on from synovial tissue [33], lung tissue [34], adipose tissue [35] and peripheral blood cells [36]. MSCs which have been derived from a wide range of tissues are remarkably different in morphology and differentiation potency. They can differentiate to a wide range of connective tissues such as bone, adipose tissue, cartilage, intervertebral disc, ligament and muscle.

MSCs showed a strong influence on the immune response process [37]. Thus, MSCs can be used in the treatment of inflammatory diseases [38]. MSCs respond to injuries or stress similar to the way native immune cells respond to apoptosis and pathogen exposure. They migrate to the site of injury and modulate the suppression of pro-inflammatory cytokines. MSCs produce a wide range of complex mediators by responding to the signals at the site of injury in order to trigger angiogenesis, regeneration, remodelling immune cell activation or restriction and cell migration. MSCs' potential to respond to the environment is ascribed to their responses to the alteration in that environment itself through regulation and translation of suitable proteins. Additionally, MSC involvement expands to organ homoeostasis, wound healing, ageing and tissue repair mechanisms.

The International Society of Cell Therapy identified the following criteria for a cell to be classified as an MSC: 1) attachment to the plastic in culture condition, 2) expression of surface markers such as CD73, CD90 and CD105 and lack of expression of CD34, CD45, HLA-DR, CD14 or CD11b, CD79a or CD19 and 3) ability to differentiate to osteoblasts, adipocytes and chondroblast in vitro [39].

The regulation of MSCs is mainly dependent on the Wnt canonical pathway [40] and TGF-superfamily pathway [41]. The Wnt pathway includes glycoproteins that trigger intracellular signalling which modulates cell proliferation and differentiation. The Wnt pathway is essential in skeletogenesis by triggering osteoblasts [40] and the inhibition

of chondrocyte development [42]. On the other hand, the TGF pathway is composed of proteins which are essential in skeletal development and modulation of chondrogenic differentiation. Moreover, several growth factors such as the epidermal growth factor (EGF), fibroblast growth factor (FGF) [43], platelet-derived growth factor (PDGF) [44] are reported as regulatory molecules of MSC differentiation.

1.4.2 Cartilage biology and disorders

The articular cartilage is a white smooth tissue which is composed of chondrocytes embedded in the ECM. It contains a low number of cells related to matrix and lacks blood and lymphatic vessels and nerves. The role of the cartilage is to provide elasticity and force-resistance on movement which is delivered by the ECM structure and composition. The cartilage propagates the pressure on joints in order to prevent joint and bone damage by excessive load force.

Chondrocytes are the only cell type residing in articular cartilage. Chondrocytes have a low level of metabolic activity and nearly no cell division, therefore show low turnover replacement. Chondrocytes are in charge of maintaining the balance between the production and degradation of the ECM in cartilage. Chondrocytes in cartilage are able to respond to a wide range of stimuli which are associated with the ECM remodelling of cartilage. These cues including cytokines and growth factors are able to modulate the differentiation and migration of the chondrocytes. The structure and function of ECM are tightly coupled and any genetic alteration in the ECM components may result in the quality of cartilage. The cartilage contains various types of proteoglycans mainly comprising of aggrecan, biglaycan, decorine, fibromodulin and lumican. These glycans are responsible for ECM hydration, buffering, bindings and force-counteract. ECM is a composite of various types of fibrous proteins. Fibrous proteins include collagens, elastin, fibronectins and laminins. They provide pliable strength, promote cell adhesion and migration, mediate chemotaxis and modulate tissue generation in the tissue [45]. Among all fibrous proteins, cartilage ECM is richest in collagen. Various types of collagen exist in each given tissue however in each given tissue only one type of collagen is normally predominantly expressed.

Lack of blood vessels, low turn over rate of chondrocytes and age-related wear and tear cause the initiation of the cartilage degradation and progression of cartilage disorders. Cartilage related disorders are the result of both mechanical and non-mechanical forces which causes cartilage degradation and development of osteoarthritis (OA). OA is the most prevalent joint disease that involves cartilage and bone erosion. OA is associated with alterations in the chondrocyte metabolism which results in increased level of

proteolytic factors, cartilage degradation and eventually dysfunctional joints. Fig.1.2 illustrates a schematic representation of a healthy vs. OA affected joint.

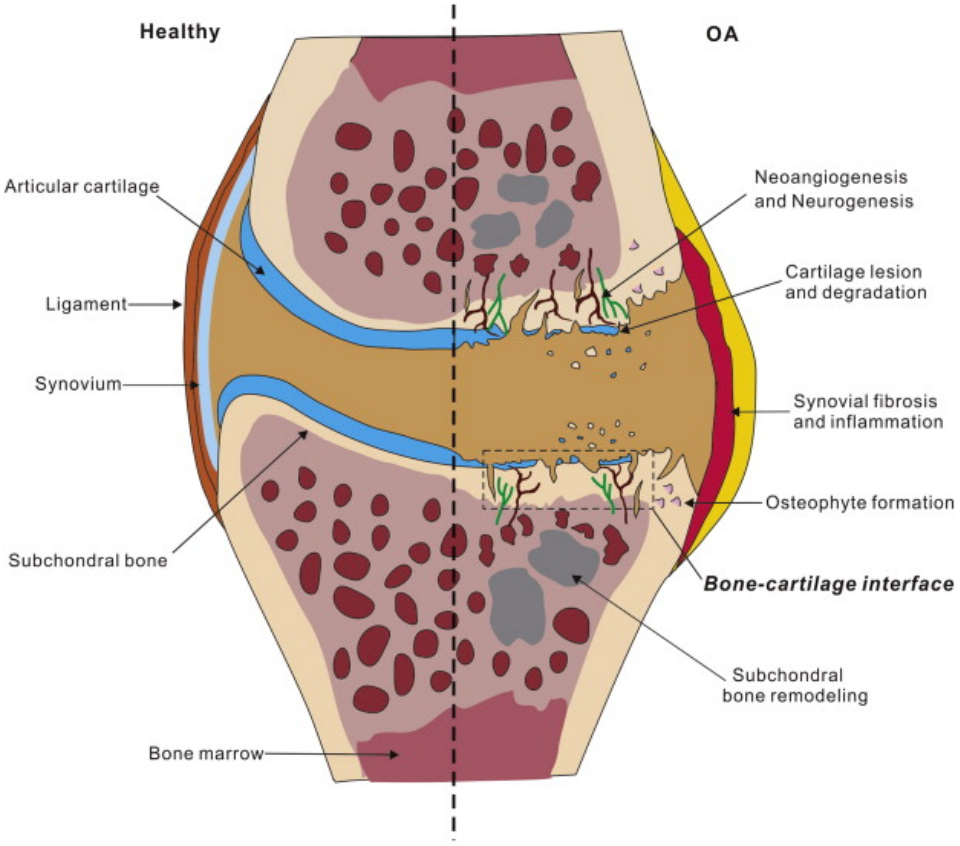


FIGURE 1.2: A schematic representation of healthy joint vs. OA.

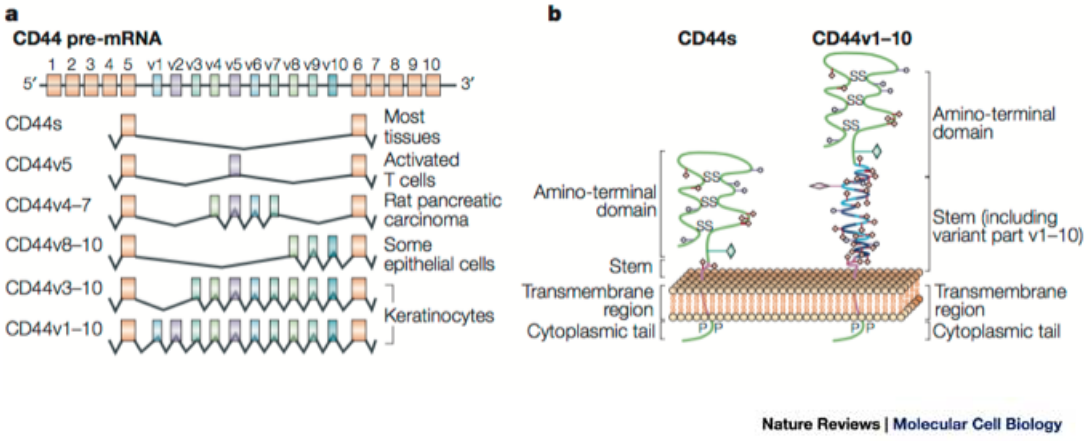
Reprinted from [46] with publisher permission.

1.4.3 Cartilage components

The selected proteins for this study are as follows: i) CD44 which is a cell surface protein, ii) Tumor Necrosis Factor-Stimulated Gene-6 Protein (TNFAIP6) a secreted protein and iii) Cartilage-Linking protein 1 (Crtl1) also known as Hyaluronan and Proteoglycan Link Protein 1 (HAPLN1), a component of the cartilage matrix. Their co-expression, co-localisation and binding ability in presence of growth stimuli and chondrogenic cues was investigated.

1.4.3.1 CD44

CD44 is a cell surface transmembrane glycoprotein which is coded by the highly conserved CD44 gene and expressed in various cell lines including chondrocytes. The



a) CD44 pre-mRNA and **b)** CD44 standards vs. CD44 variants. Reprinted from [47] with publisher permission.

FIGURE 1.3: A schematic representation of CD44.

CD44 gene is located on the plus strand of chromosome 11 and starts from 35138870 to 35232402. CD44 is composed of 20 exons and is expressed in a variety of tissues with a diverse range of functionalities mainly due to its alternative splicing and post-translational modifications. MSCs generally lack the variant exons while differentiated ectodermal cells comprise all the variants. The standard and shortest isoform of CD44 with 80 kDa molecular weight includes exons 1-5 forming a globular domain and exons 16-20. This standrad isoform is expressed in most cell types. Standard CD44 has an N-terminal signal sequence coded by exon 1, the 'link' domain for binding to Hyaluronic Acid (HA) coded by exon 2 and 3, and a stem region composed of exons 4,5,16,17. The hydrophobic transmembrane segment and a cysteine residue are formed by exon 18. The cytoplasmic region of CD44 which incorporates motifs with the capability of signal transduction is constructed by exon 20.

Even though a wide distribution of cell types express standard CD44 the insertion of variant exons V1-V10 gives rise to CD44 variations which are observed in cancer cells. The variant exons are inserted in the stem structure and elongate the stem segment. Certain CD44 splicing variants are recognised to be involved in metastasis of cancer cells and have provoked a large number of studies with a focus on diagnostic and prognostic tests of CD44 standard and CD44 variants in the various cancer type. The majority of these studies provide uniform evidence that indicate the abnormal over expression of CD44 splicing variants. However, in the case of prostate cancer and neuroblastoma the expression level of CD44 does not correlate with the metastasis and shows inhibitory effects on tumour invasion. In the majority of cancer types the genetic mutation of CD44 itself is not observed, however, genes which are associated with carcinogenesis govern CD44 expression. Mitogenic signalling such as the Ras-MAP kinase pathway influences

the splicing variants of CD44 [48, 49]. Additionally, the chromatin remodelling complex of SWI/SNF is mutated in various types of cancer which leads to the lack of CD44 transcription [50, 51]. Therefore unusual expression of CD44 is dependent on tumour promoting genetic alteration. CD44 loss of function studies suggest its role in tumour formation, immune responses, bacterial infection and haematopoiesis [47].

The CD44 protein family has crucial roles in cellular behaviour and any abnormality in their expression pattern, splicing variants or post-translational modification leads to pathogenic phenotypes. CD44 is involved in a wide range of cellular processes such as organ development, neuronal axon guidance, immune responses, adhesion, angiogenesis, inflammation and tumour development and haematopoiesis [47].

During inflammation, the level of CD44 is elevated in hematopoietic cells. It binds to T-cells and triggers the migration of T-cells to the inflammatory sites. The ability of CD44 to bind to collagen, fibronectin and osteopontin causes T-cells to adhere and be retained in the inflammation sites.

Alternative splicing and post-translational modification give rise to diverse structures of CD44 which consequently influence its functionality and mechanism of action. Three groups of molecular functions are associated with the CD44 protein family: ligand binding function, co-receptor function and plasma membrane and actin cytoskeleton linkage.

As a cell surface receptor CD44 can interact with a number of ECM components such as hyaluronan, collagen, laminin, fibronectin as well as other glycosaminoglycans (GAGs). The cell type specific post-translational modification seems to influence the binding affinity of CD44 to GAGs. In addition to soluble extracellular components and ECM components, CD44 is able to act as cell-surface receptor protein that cross-links enzymes and their substrates.

The binding affinity between HA and CD44 is influenced by several factors such as mitogenic cues, glycosylation of extracellular and phosphorylation of certain serine residues in the cytoplasmic tail of CD44. Its binding affinity is essential for cell migration and leukocyte flow.

The extracellular domain of CD44 is prone to proteolytic cleavage in various cancer types which suggests the existence of other mechanisms that regulate CD44-hyaluronan binding. The deterrent of CD44 cleavage resulted in suppression of tumour cell migration. CD44 acts as a platform that facilitates binding between enzymes and their substrates. An example of such interaction is Matrix Metalloproteinase (MMPs) and growth factors. Interaction of CD44 and MMP9 in mouse mammary cell lines and human melanoma cell lines mediated the degradation of collagen IV and promoted tumour

invasion [52]. In addition, MMP9 bound to CD44 could activate inactive transforming growth factor (TGF-β) and consequently could promote neovascularization. The same study concluded that the formation of metastasis is triggered by CD44 dependent inhibition of apoptosis, collagen IV degradation and neovascularization.

Additionally, CD44 acts as co-receptor and regulates the signalling cascade of the receptor tyrosine kinases, in particular, the Met and ERBB family [47]. Met is a tyrosine kinase receptor which is essential for embryonic development and wound healing and ERBB family contains receptor tyrosine kinases which are structurally related to the epidermal growth factor receptor (EGFR). This ability can explain the role of CD44 in the growth and metastasis of cancer cells. The majority of these receptors and their ligands act in the form of complexes. One possible role of CD44 is the coordination of the establishment of these complexes in order to secure a flawless signal transmission. The activation of tyrosine kinases is a complex process in which multiple co-receptors are involved and cellular adhesion proteins can act as co-receptors. The CD44 isoforms co-receptor function describe the cell migration and tumour metastasis.

CD44 binds to cell surface receptors such as ERM protein family (ezrin, radixin and moesinand) and merlin whose interaction with underlying cytoskeleton ensures efficient signalling for cellular migration [47]. ERM mediates signal transduction between tyrosine kinases and the Ras pathway while activated merlin suppresses Ras and its downstream components. In most cancer cell lines, excluding prostate cancer and neuroblastoma, CD44 stimulates metastasis. These dual mechanisms of action indicate a switch like role for CD44. Thus CD44 can translate cellular signals to both invasive and non-invasive processes which requires a tight control in order to maintain the optimal balance. The imbalance between either of these states influences the tumour status. Despite the complexity of CD44 action due to its dual switch functionality, targeting CD44 serves as a propitious therapeutic approach.

1.4.3.2 TNF-stimulated gene 6 protein (TSG-6)

Tumour necrosis factor-inducible gene 6 protein is a secreted protein which is encoded by the TNFAIP6 gene. It is located on chromosome 2q23.3 and its molecular weight is around 30 kDa and has a conserved sequence among mammals (94% sequence similarity) [53]. It is expressed in various cell types including chondrocytes, synoviocytes, vascular smooth muscle cells in response to wide range of growth factors. TSG-6 is not continuously present in cells; however, a wide range of factors can stimulate its expression including TNF-α and IL-1 which are pro-inflammatory cytokines thus it is an indication

of TSG-6 involvement in the inflammatory process and its association with inflammatory diseases. TSG-6 has been detected in the sera of patients with bacterial sepsis [54, 55], and joint tissues and synovial fluids of patients with different types of arthritis [56]. TSG-6 is composed of a hyaluronan-binding Link domain, therefore ECM stability and cell migration is the major cellular process that TSG-6 is involved in. It interacts with multiple ECM components including GAGs, HA [57], chondroitin-4-sulphate (C4S) [58, 59], aggrecan and serine protease inhibitor inter-α-inhibitor(IαI) [60, 61]. In addition to the Link domain, TSG-6 contains a CUB module (complement C1r/C1s, Uegf, Bmp1). CUB is around 110 residues and are found in extracellular and plasma membrane-associated proteins. These two domains are identified on residue 37-128 and 129-250 respectively [57, 62]. TSG6 binds to HA via its link domain. Similarly, the link domain is its binding site to C4S and G1 domain of aggrecan [59]. However, CUB domain provides a binding site to heparin.

TSG-6 makes a covalent bond with Inter-alpha-inhibitor (IαI) which is a family member of the serine protease inhibitor. IαI is an HA binding protein reported for the first time by Sandson et al.[63] who detected the HA-IαI complex in synovial fluids. HA is an important component of connective tissues and is involved in intracellular signalling. TSG-6 binds to both IαI and HA, therefore, it is hypothesised that it can modulate the complex formation of HA-IαI and plays a vital role in ECM assembly and remodelling. IαI is the inhibitor of a wide range of proteases including plasmin. Wisniewski et al. [61] showed that the moderate anti-plasmin function of IαI is enhanced by TSG-6. In a study conducted by Getting et al. [64] recombinant Link_TSG-6 enhanced the IαI anti-plasmin role even though TSG-6 does not bind to IαI covalently. This fact suggests that covalent binding is not necessary for modulating anti-plasmin activities. Moreover the TSG-6_IαI complex which has been characterised is lacking the bikunin chain, thus it would not demonstrate anti-protease activity. Based on this evidence it has been assumed that TSG-6 either influences the formation of the HA_IαI complex in ECM or non-covalently binds to IαI and augments the anti-plasmin activity of IαI in the protease network.

TSG-6 is associated with arthritis and its overexpression was detected in synovial fluids and sera of patients with different types of arthritis. Interleukin-1 (IL-1) and tumor necrosis factor (TNF) modulate the expression of TSG-6 in articular synoviocytes [56]. IL-1, TNF, PDGF and TGF-β regulate its expression in articular chondrocytes [65, 66]. TSG-6 is also expressed in synovium and cartilage of OA and rheumatoid arthritis (RA) joints [67]. A decreased level of aggrecan and an increased level of TSG-6 in ECM surrounding the OA-like lesion indicate that TSG-6 might compete with aggrecan in binding to HA molecules [68]. Moreover, TSG-6 was detected in the chondrocyte pericellular matrix of STR/ort mice preceding the development of OA lesions which is

an indicator of disease initiation. Midrescu et al [69] showed the local expression of TSG-6 in the arthritic joint restricts inflammation and thus prevents cartilage destruction and bone erosion in collagen-induced arthritis (CIA) mice.

TSG-6 also has a chondroprotective function. Bardos et al. [70] injected rm-TSG-6 into severely inflamed joints of mice with antigen-induced arthritis (AIA) and showed no aggrecan fragments are observed which suggested MMPs that degrade the cartilage were suppressed. In transgenic mice which have been developed by Glant et al. [71] where murine TSG-6 is expressed in cartilage, the induction of AIA led to acute joint inflammation but the cartilage stayed undamaged after 1 week. Also, MMP-generated fragments of aggrecan were decreased. At week 4-5 the cartilage in transgenic mice was healed and no local inflammation was observed. This evidence suggested that the anti-plasmin effect of TSG-6_IαI is responsible for such a chondroprotective effect and suppression if cartilage MMPs with matrix degradation function (MMPs 1,2,3,9 and 14). Bardos et al. also observed that induced TSG-6 increased at inflammation sites which resulted in capturing IαI from serum. In connection with these results both complexes of TSG-6_HA and TSG-6_IαI have been detected in synovial fluids of patients with arthritis [56, 72].

1.4.3.3 Hyaluronan and proteoglycan link protein 1 (HAPLN1)

Link protein or Hyaluronan and proteoglycan link protein 1 (HAPLN1) residing on Chromosome 5q14.3 is best known for stabilising interactions between HA and proteoglycans such as aggrecan [73]. The HAPLN1 sequence is conserved across human, rat, chicken, pig and bovine. HAPLN1 consists of 3 domains which are built from a disulfide bounded loop. The amino terminus of the protein forms an immunoglobulin-like (Ig-like) motif which is involved in interaction with aggrecan [74]. The Ig-like motif is followed by two tandem repeat motifs (TR1 and TR2) which enable the HA binding [75, 76]. The Link protein was originally detected in cartilage, however, it is also expressed in brain [77], aorta [78], mesonephros of chick embryo [79]. Aggrecan is not expressed in these tissues, however, other proteoglycans such as versican and neurocan which bind to HA are expressed in the aforementioned tissues.

Link protein serves as a coating for the HA molecule in order to prevent its cleavage by hyaluronidases or free radicals [80] and Link protein also delays the aggregation process during aggrecan formation [81]. The delay in the aggregation will prevent aggrecan retention in the cell and allows it to circulate to the ECM prior to aggregation. Newly secreted aggrecan is not capable of binding to HA. However, later on by a process which

involves the formation of disulphide bonds between aggrecan's G1 domain and Link protein, aggrecan will be able to bind to HA molecules [82].

The regulation of HAPLN1 is under the tight control of Sox9 similar to other cartilage matrix genes including COL2A1, COL9A1, COL11A2 and aggrecan. Studies have shown that the expression of HAPLN1 in developmental bone is governed solely by the level of SOX9 [83]. However, the deletion of SOX9 in the heart valve did not entirely eliminate HAPLN1 which indicates HAPLN1 is not entirely sustained by SOX9 [84]. Mef2c is the co-factor of SOX9 in modulating the chondrocytes, hypertrophy [85, 86]. Cartilage-specific SOX9 deletion in combination with Mef2c or independent of Mef2c stalled the chondrocyte hypertrophy and resulted in shortened bones. Hence, Mef2c might be another regulator of HAPLN1.

1.5 Aims and objectives

This thesis is set out to investigate a novel approach in the prediction of PPIs. More precisely protein features which could be employed in computational models were used for the prediction of PPIs. In addition, the application of Linked Data principles for an integrative approach was used in exploring and inferring interactions between proteins from existing data. The predictive model and integrative platform were were complemented selectively validating novel PPIs with experimental methods.

The first research question focuses on the genomic location of the genes and their relevance to the protein functionality both based on the linear positioning of the genes and their 3D positioning with regards to the chromosomal structure in the nucleus. The linear or 2D genomic location is defined as the number of bases between the *end* codon of the first gene and the *start* codon of the second gene. However, this measurement applies only to the genes located on the same chromosome and is not applicable to the genes which are located on two different chromosomes. This research question was elaborated with more detailed analytic in Chapter 3. In this thesis, a linear distance model was translated to a 3D distance based on the genomic locations of pairs of proteins. A scoring function was introduced which ranks the 3D genomic feature between protein pairs. The relation of genomic location of the genes with the position of their proteins in the PPIs network were investigated and the functionally related genomic segments were identified based on the genomic location of protein partners.

In the second research question, the predictive power of genomic location as a predictive feature was investigated with the state-of-art machine learning approaches in combination with other descriptive features. Results of this comparison are reported in Chapter 3.

The third objective focuses on the development of a PPI integrative framework based on Linked Data principles. The proposed framework integrates the available genomic data including PPIs, gene expression, domain composition, GO similarity and genomic location of proteins. In this framework, a domain specific model was introduced which connects PPI related concepts by a set of vocabularies. This framework is presented in Chapter 4 where multiple data sources were semantically integrated.

The final objective of this thesis is to validate the proposed *in silico* findings with *in vitro* methods. Thus, a human protein, CD44, which is involved in cartilage formation was selected and its network of interactions was validated based on the novel interactions identified in the proposed framework. The experimental validation component of this thesis focuses on laboratory methods, materials and results which is presented in Chapter 5.

The following chapter of this thesis will demonstrate an overview on the state-of-the-art computational models and provide a comparison study of the performance of these models while using various protein features including sequence, structures and function.

Chapter 2

Related Works

Predictions of protein-protein interactions (PPIs) is considered a two-class classification problem. Generally, these two classes are the interacting protein dataset (positive gold standard) and the non-interacting dataset (negative gold standard). The positive dataset is retrieved from various publicly available databases such as BioGRID [15], HRPD [17], MINT [18], IntAct [19]. However, collecting the negative dataset (the non-interacting protein pairs) is not similarly straightforward due to the incompleteness of PPI networks for human and other model organisms. The two most frequently used approaches for generating non-interacting datasets are randomisation and the identification of non-co-localised proteins. For randomisation, the studies have used random pairs of PPIs based on the proteins available in the interacting dataset (the positive set). Apart from randomisation, other studies have generated pairs of proteins that vary in their cellular localisation making use of the fact that proteins can only interact with each other if they share their location. Thus if they reside in different cellular compartments then there is only a limited likelihood that they interact. Both interacting and non-interacting datasets are transformed into a set of features in order to be used by the classifier (typically any machine learning-based algorithms). Regardless of the type of classifier, the feature set for each protein and protein pair in the positive and negative dataset is generated based on the chosen descriptive attributes for each protein.

The following sections describe commonly used protein features and classifiers which have been applied to address the PPI prediction problem (without claiming to cover the full scope of the extensive research works done). Following this part, a comparison of performances of these related works which used the presented algorithms and features is provided. Also, this chapter includes a brief overview of biomedical data standardisation initiatives which employed Linked Data concepts as a mean to integrate multiple data sources.

2.1 Modelling of PPI predictions

2.1.1 Descriptive features

Based on the nature of the data used in the prediction of PPIs, features can be classified as *sequence based* and *concept based*. Either approach transforms protein characteristics into a vector representation of the protein's features. Sequence-based methods achieve better prediction accuracy in comparison to conceptual features due to the fact that all the conceptual features might not be available or characterised for all proteins. As a result, the missing information introduces incompleteness to the feature vectors and consequently affects the result performance of the model. The following section describes the most widely used features for the classification of PPIs.

2.1.1.1 Sequence-based features

The amino acid sequence of a protein contains relevant information for its 3D structure which specifies its interaction sites for the interacting protein partners. Recently, methods that exploit PPI patterns directly derived from the proteins' amino acid sequences attracted researchers' interest [27, 30]. The development of prediction methods solely based on sequence information and on the physio-chemical properties of amino acids [87], the conjoint trait [88] and the effect of their neighbouring conditions in continuous and discontinuous sequence segments [29] showed promising results for the identification of novel PPIs [89]. The physiochemical parameters of amino acids play an important role in maintaining the stability of PPIs [87]. As a result, methods which incorporate these features show higher performance in the prediction of PPIs, in particular a higher accuracy rate. In these studies, protein sequences are described as a set of numerical values which characterise hydrophobicity [90], the volume of side chains of amino acids [91], polarity [92], polarizability [93], the solvent-accessible surface area [94] and the net charge index of side chains [95] for each amino acid in the protein sequence. As an example, a protein sequence with the length of 100 amino acids will produce a vector of 100 x 6 = 600 values which introduces a high level of dimensionality into the data. On the other hand, hydrophobic and electrostatic bonds are more frequently observed in PPIs therefore, Shen et al. [88] clustered the 20 amino acids into 7 groups based on their dipoles and the volume of the side chains. This approach reduces the dimension of the data drastically which is favourable when machine learning algorithms are applied for the classification tasks. As a consequence, this approach has been adopted by several other studies [89, 96].

Interaction sites in proteins are formed by the combination of several amino acids in the protein sequence, therefore the feature of conjoint trait has been introduced by Shen et al. A conjoint trait is defined as the sequence of three amino acids in a row in the protein sequence while each amino acid can be categorised into one of seven clusters. This way the total number of possible traits results to 7x7x7= 343 variants. Protein sequences vary in length but if they are transformed to the frequency of observation of each possible trait combination, then the feature vector for all proteins with any sequence length will be of equal size. The main advantage of using the conjoint trait feature is reduction in the dimensionality of the data. However, conjoint trait does not account for the interaction information coded in the neighbouring amino acids therefore Gua et al.[29] proposed auto co-variance feature on the continuous and discontinuous segments of the protein sequence. Auto co-variance takes into account the interactions between distantly located amino acids. Hence, this feature accounts for the neighbouring effect of amino acids. Therefore, auto co-variance captures the detail of sequence information as well as decreasing the feature space.

2.1.1.2 Concept based features

Concept based features form a combination from a set of features which are not derived from the amino acid combination of the protein sequence. Generally, the PPIs from the gold standard dataset, i.e. from the core PPI data source (a database), are transformed to a vector which contains the selected combination of these conceptual features producing the concept based feature set. The main drawback of representing the interactions as a set of concept based features is the incompleteness of the underlying data source for these features. Since proteins are typically annotated either through features from experimental methods or from computational methods, such annotations are often missing in the primary data source for a large portion of relevant proteins. Thus, once the PPIs have been transformed into the feature vectors the missing annotations introduce incompleteness to the overall data representation. As a consequence, the prediction outcomes of the models are quite restricted. The following list gives a brief description of a selection of these features:

Sequence similarity based features assume homologous proteins share sequence similarity, leading to the similarity in their structure and functionality [97].

Gene neighbouring assumes higher functional similarity between genes which are located in close proximity on the genome. Studies have shown [98–100] that this is predominant in prokaryotes while there is no strong evidence for such phenomenon in higher organisms like humans.

Phylogenetic similarity is measured by the co-existence of a pair of proteins across multiple organisms based on the assumption that functionally related genes evolve together over time [101–103].

Gene-expression patterns have been used in several studies and assume that the two interacting proteins are expected to show similar expression patterns [104]. However, it is important to refrain from the inverse conclusion, i.e. two genes with similar expression patterns may not interact at the protein level.

GO-based similarity measures the similarity of concepts from GO annotations shared between two proteins [105]. The GO annotation comprises the *cellular localisation*, *molecular function* and *biological process* for a protein. Two proteins are only able to interact if they are located in the similar cellular component. Therefore, two proteins with similar cellular localisation annotations have a higher likelihood to interact in contrast to those located in different compartments. Similarly, proteins with shared functional annotations or with involvement in similar processes are more likely to interact with each other.

Protein domain composition is another reliable predictive feature. Domains are structural components of proteins enabling specific functions and each protein is composed of one or several of such domains. Interacting proteins carry out their functions by combining their domains. Hence, interactions between protein domains are indicative for interactions between proteins.

2.1.2 Prediction algorithms

This section gives a brief overview of a selected number of algorithms which have been successfully applied for PPI prediction research problems exploiting the aforementioned protein features.

Naïve Bayes (NB) is a probabilistic classifier based on the Bayes theorem which assumes independence between variables. NB uses the maximum likelihood method to estimate the required parameters (based on a small set of training data). Despite its simplicity this method is very efficient in providing a solution to real life prediction problems including the prediction of PPIs from feature based data. The independence assumption between features turns NB into an efficient algorithm which produces similar performance in comparison to other more complex approaches. By contrast, features in real life scenarios are often related to each other and thus dependent, and therefore the conditional independence assumption may not entirely serve well for real life problems like PPIs. Saha et al. employed NB in combination with support vector machine

(SVM), random forest (RF) and decision trees in forming an ensemble learning algorithm which obtained 90% accuracy. PrePPI [106] applies NB to a set of structural and non-structural features. PrePPI identifies structural features of proteins as the main contributor to the underlying algorithm. Tree-augmented Naïve Bayes (TAN) is an extension to a previously build predictive PPI model based on NB by Lin et al.[107]. Unlike the original Bayes theory, TAN allows for the dependency between the feature set comprising of gene expression, gene ontology and orthologous scores. They reported that TAN achieved the ROC area of 0.8 which is higher than 0.75 and 0.78 obtained by unrestricted Bayesian and NB, respectively. Phylogenetic profiles of protein pairs serve as features of the NB model in order to predict the functional relevance of the pairs [108]. Co-evolutionary divergence similarity between pairs of proteins in 14 model organisms were used as predictive features for the NB algorithm [109]. In another study Najafabadi et al.[110] applied NB to the codon composition of the protein sequences and was able to identify functional and physical protein partners and eventualy, served as an effective predictive model.

Support vector machines (SVMs) form a supervised learning algorithm which separates the two classes of data by drawing complex hyperplanes between sets of data for the classes. These hyperplanes maximise the margin to the nearest data points belonging to each class. Often, the correlation between features of data points are not linearly separable and special kernel functions are required in order to classify highly complex problems. Using kernel functions in SVMs requires parameter estimation and optimisation solutions. Therefore the prediction power of SVMs is influenced by the parameter estimation and is considered to form the main drawback of SVMs.

Despite their complexity and their computationally expensive performance, SVMs have been successfully applied to several classification problems in computational biology including PPI prediction problems. Mahmoudian et al. [111] published a model which is called Support Vector Regression (SVR) and is a variant of SVMs. Since regression is applied to analyse continues variables and PPI labelling is a two class categorical problem, the SVR had to transform its real value into a label. The authors addressed this problem by introducing a mapping point which is a cut off point that maps the pairs of proteins to either of the interacting or non-interacting classes.

SVMs are computationally expensive in particular if the predictor features are based on the amino acid composition of the protein sequences. A distributed and parallel computing infrastructure can be employed to improve the training time. You et al.[**?**] proposed a MapReduce-based parallel SVMs which demonstrates to solve the high dimensionality of the data as well as maintaining a high level of accuracy. An SVM model was built based on genetic codons of *M.tuberculosis* along with the ortholog protein

interactions from 14 different species [112] to infer PPIs. Similarly, sequence-based protein partners search (SPPS) is a method which employs probability-based SVMs to the conjoint trait of amino acid sequence and is able to predict both direct and indirect interactions between pairs of proteins [113]. Despite their complexity and computational exhaustion several lines of evidences aforesaid state that SVMs are the best candidate in sequence based classification in which the feature set exposes complexity through its high dimensionality.

Random Forest(RF) is a classifier built up from several decision trees where each tree is constructed from random and independent features in the feature space. In order to classify a new instance, the input node is sent to the trees and each tree votes for a label. Ultimately the final class label is selected based on the majority of votes of the trees in the forest. RF is based on the *divide-and-conquer* principle and assumes that contributions to the output from each decision tree improves the prediction performance. Each decision tree in the forest is a *weak learner* and the ensemble of trees forms a *strong leaner*. The top-down principle (i.e. the divide and conquer) of this algorithm allows the *unseen* instances of the data to traverse through the forest to the lowest level of the trees (down to the terminal nodes) while the data is narrowed down as part of the same procedure. In each level of the tree, a random variable and its value is determined in a way that the variable optimises the split of data for the next level. The main advantages of RF in comparison to a decision tree are its ability to avoid over-fitting and thus makes it suitable for small sample size datasets as well as the ascription of missing data.

RF has been widely used in PPI prediction studies imputable to its unexcelled accuracy and efficiency in running on datasets with missing data. LocFuse [105] classifies PPIs based on eight features with a special focus on cellular localisation of proteins by application of several classifiers including RF. Mohamed et al. [114] improve the efficiency and predictive power of RF by introducing active learning to the algorithm.

k-nearest neighbours (kNN) is a classification method which labels an unlabelled data point based on majority vote of its k nearest labelled data points in the feature space making use of a distance function. kNN is a *non-parametric* method which does not make any assumption regarding the distribution of the data. It is also a *lazy* learner which means that the training is fast because it postpones the generalisation but the testing is lazy and it keeps all the training dataset in memory, therefore it requires longer time and more memory.

K is a user-defined value which requires optimisation in order to improve the classification outcome and it is generally an odd value for a two class problem. The optimal estimation of k is an essential step in order to improve the accuracy of the classifier. kNN assumes all features have similar effects on the distance between the neighbouring

data points and the unlabelled instance. Thus it is not able to differentiate between irrelevant features. Therefore any increase in the feature space dimensionality will reduce the classification accuracy. However, an extension of kNN called *weighted kNN* addressed this problem by assigning weights to the neighbouring labelled nodes based on their distance to the *unlabelled* node. In other words, the closer the neighbouring node is, the higher is its vote in contrast to the neighbours that are further away.

kNN has previously been applied to multiple PPI prediction studies, however any increase in the size of the dataset or feature sets directly increased the computational cost as well. Hence it is crucial to optimise the scale of the feature set. Coupling kNN and a genetic algorithm showed a promising outcome in the characterisation of protein phenotypes based on PPI network topology and protein sequence information [115]. In a similar study which aimed to predict the protein functionality Lan et al. [116] proposed a multi-source kNN (MS-kNN). MS-kNN predicts protein functions based on the proteins' neighbours' characteristic, in particular, their sequence and gene expression similarities. However in a study [117] which has used conjoint traits of the protein sequences as the feature set, the authors have reported higher predictive power for kNN in comparison to SVM as the state of the art.

Up to now, the most relevant biological attributes of proteins for the classification of PPIs and the algorithms which have been used in this regard were described. As part of the background research, a comparison study has been performed on previously published studies which used the above features and algorithms either individually or in combination with others in order to explore the impact of the feature selection on the outcome of the different known and tested approaches. The following section constitutes the outcome of this comparison study.

2.1.3 Evaluation and comparison of performances for state of the art approaches

In order to assess the prediction performance of proposed methods, the following parameters have generally been measured: accuracy (ACC), sensitivity (Sens) or recall , specificity (Spec) and F1 score which is the harmonic average of precision and recall. These parameters are defined as follow:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

$$Sen = \frac{TP}{TP + FN} \tag{2.2}$$

$$Spec = \frac{TN}{TN + FP} \tag{2.3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{2.4}$$

Where true positive (TP) is the number of correctly classified interacting PPIs, false positive (FP) is the number of falsely labelled non-interacting PPIs as interacting PPIs, true negative (TN) is the number of truly non-interacting pairs and false negative (FN) is the number of interacting pairs which have been classified as non-interacting pairs.

Table 2.1 contains a summary of the performance metrics of the most representative methods. It is noteworthy that all published methods have not been using the same set of evaluation metrics, therefore some of the values in Table 2.1 are missing. However, in most cases the reported parameters are interchangeable.

Based on the data which is shown in Table 2.1, it can be concluded that the best performances have been achieved when sequence based features were used as predictor features. The highest accuracy was reported by Liu et al. [112] which came close to 95% by the use of SVM on the composition of codons of the sequence in combination with interlog data. The second highest accuracy was attained in a study by Xia et al. [87] in which surprisingly a decision tree algorithm performed equally well as the SVM base models using auto-correlation features of the protein sequences. They reported 93.5% accuracy while similar features were used in another study by Yoa et al. [119] and the area under the curve (AUC) reached 0.97. The higher the AUC score and the closer it is to the value 1.0, the better is the performance of the approach. However, in this case, an exact comparison is not feasible since two papers reported different metrics for their model performance.

While observations from Table 2.1 suggest that an SVM in general performs well on sequence based features, the complexity of model organisms consequently extend the number of PPI available on model organisms influence the performance of the algorithm. Focusing on the results that have been reported by Liu et al.[113] using five different model organisms with SVMs and conjoint trait features, the performance of the model varies between 83% to 91% for C.elegance and yeast, respectively. Since conjoint trait features are solely based on the sequence information which is available for all five model organisms, the variation in the accuracy can be attributed to the sequence diversity for

TABLE 2.1: Summary of performance of state of the art models.

Paper	Organism	Interactions	Algorithm	Ensemble	Features	AUC	Sens	Spec	F-Score	Accuracy
Saha et al.[118]	Hm, Yt	57576, 190377	SVM, RF, DT, NB	✓	GO, PD	0.64	0.1	0.67	0.25	0.77
Lin et al.[108]	Yt	1073845	NB	-	Phylogenetic profiling, Sequence base	0.67	0.001	-	-	-
Liu et al.[109]	Hm	36867	NB	-	Co-evolution	0.63	-	-	-	-
Mahmoudian et al.[111]	Hm	10516	SVM Regression	-	KUPS dataset	-	0.64	0.8	0.71	0.74
Yoa et al.[119]	Hm	36630	SVM	-	Autocorrelation descriptor (sequence base)	0.97	0.9	0.9	-	-
You et al.[120]	Yt	11188	SVM	-	Sequence base	0.97	0.9	0.91	-	0.91
	Hp	2916		-		-	0.83	0.86	-	0.84
Liu et al.[112]	Mb	41940	SVM	-	Codon (sequence base), Interologs	0.87	0.89	0.85	-	0.95
Liu et al.[113]	Hm	39191	SVM	-	Conjoint trait (sequence base)	-	0.82	0.97	-	0.9
	Ce	4973		-		-	0.77	0.9	-	0.83
	Dm	22482		-		-	0.8	0.95	-	0.88
	Yt	25064		-		-	0.85	0.97	-	0.91
	Mm	1225		-		-	0.8	0.88	-	0.84
Xia et al.[87]	Hp	2916	DT	✓	Autocorrelatin descriptor (sequence base)	-	0.9	0.96	-	0.93
	Yt	11188		-		-	0.88	0.89	-	0.88
Zahiri et al.[105]	Hm	24350	Rf,NB, MLP, RBF	✓	PTM, Codon Usage, Tissue information, PSSM , GO	0.85	0.7	0.81	0.76	0.77
Hayashida et al.[121]	Hm	898	CRF	-	PD	0.91	-	-	-	-
	Dm	588		-		0.78	-	-	-	-
	Ce	500		-		0.88	-	-	-	-

Abbreviations: Hm: human, **Yt**: Yeast, **Ec**: E.coli, **Hp**: Helicobacter pylori, **Pf**: falciparum, **Ce**: C.elegance, **Dm**: D.melanogaster, **Mm**: M.musculus, **Mb**: Microbacterium. **AUC**: Area under the curve. **Sens**: Sensitivity. **Spec**: Specificity. **MCC**: Mathew correlation coefficient. **SVM**: Support vector machine, **RF**: Random forest, **DT**: Decision tree, **NB**: Naïve Bayes. **CRF**: Conditional random forest. **GO**: Gene Ontology. **PTM**: Post translational modification. **PSSM**: Position specific scoring matrix. **PD**: Protein domain.

the different organisms. Thus this evidence reduces the expected potential from conjoint trait features as the best sequence based predictive features. The very low performance for sequence based models was identified with an AUC score of 0.67 in which NB was employed by phylogenetic profiling features on a yeast dataset [108].

In contrast to sequence based features, the concept-based features performed poorly. The highest achieved performance resulted to an AUC of 0.91 (by Hayashida et al.[121]) while using domain composition of proteins with a conditional random forest on a human dataset. Domain composition refers to the annotation of segments of protein sequences, therefore it is not entirely independent of the protein sequence which indirectly determines the functionality of a protein. Hence algorithms that use domain composition information perform better in comparison to those which use other sources of concept based features. Similarly, studies which used the gene ontology and domain composition [118] or post-translational modifications (PTM), a position-specific scoring matrix (PSSM) and tissue information [105] reported only 77% accuracy despite the application of ensemble algorithms which is expected to improve the model performance.

2.2 Standardisation of biomedical data

Studying phenotype-genotype relations and dependencies has raised the attention of a multitude of researchers, thus several attempts have been made in order to establish data sources which deliver the genotype-phenotype relations. One such example is DisGen which is the most extensive database on gene-disease association. DisGen-RDF [122] is the RDF version of DisGen provided by Rosinach et al. By re-using the existing ontologies the DisGen-RDF transforms all the concepts which are embedded in the original dataset of DisGen – including the genes, diseases, their associations, SNPs and their association scores. The DisGen-RDF ontology is developed in alignment with other biomedical ontologies thus it allows users and domain experts to federate the DisGen-RDF with other sources of biomedical data including gene expression, drug associations, networks and pathways and numerous other biomedical data sources.

The protein functionality not only depends on the protein's structure but also on post-translational modifications (PTM) that influence the protein's functionality. One such modification is the addition of glycans or sugar chains to the sequence of a protein. Understanding cellular glycomics is essential in discerning cellular processes and their malfunctioning which eventually impairs the healthy state of the cell. Similar to most genomics and proteomics data sources, glycan data sources lack a unique data format and data interoperability. Barrentt et al. [123] developed the Glycome Analytic Platform (GAP) which aims at the integration of available data sources related to the

glycans. GAP is a suite of tools which are embedded in the Galaxy bioinformatics platform. Data sources which have been integrated including the Consortium for Functional Glycomics (CFG) data source, comprise data concerning glycan profiling, glycan array, gene-microarray and mouse phenotyping. In order to comply with other widely used data sources GAP provides access to the KEGG database as well. Glycomis data encloses a wide range of data sources including those which are directly related to the glycan itself (e.g. structures) or those which are associated with the experiments that have been carried out on these glycans (e.g. publications). Therefore, developing a standard format is not a trivial task. In order to understand all these complexities, Ranzinger et al. introduced the GlycoRDF ontology [124]. The GlycoRDF ontology introduces a structure that is required in a wide range of glycomics data sources as well as re-using the widely used ontologies such as UniProt core, the Bibliographic Ontology, Initiative (HUPO-PSI) and Mass Spectrometry Ontology (PSI-MS) [125].

Drug development studies involve extensive toxicologial evaluations of every single chemical compound, which could be a drug candidate. However, these toxicology studies are inevitably coupled with other *'-omics'* studies mainly in genomics. Toxicogenomics research analyses the relation of gene expression of target model organisms for specific biomarkers. Toxygates [126] is an initiative that attempts to semantically link the toxicology of compounds, pathways and microarray expression profiles of biomarkers and compounds interactions. Toxygates is freely available to users; however, its original project, The Japanese Toxicogenomics Project (TGP) [127], is not open for public use.

Similar to toxicogenomics which brings multiple *'-omics'* results together in order to address research questions regarding compound toxicology, the IntergromeDB [128] also amalgamates a diverse range of resources in order to facilitate the research concerning transcriptional regulatory elements. IntegromeDB is a graph based system which allows users to use their own data as well as search and capture the relevant data sources. In contrast to comparable systems (e.g. ONDEX [129], BIOZON[1] and BNDB [130]), IntegromeDB takes into account the sequence information. The integration of sequence annotations is the advantage of IntegromeDB over other comparable systems.

Additionally, VCF2RDF [131] is a similar approach for transforming the already existing data format to RDF. The Variant Call Format (VCF) was originated from the 1000 Genomes Project and contains polymorphism data. VCF2RDF creates links to the isoforms, and thus provides a unique identifier to each isoform. This enables better communication between models which integrate isoforms with other data sources.

Komiyama et al. proposed a model for the prediction of the protein-ligand binding affinity [132]. They made use of semantic web technologies in order to link the data

[1] http://biozon.org/

sources which have been used in their proposed model. Their model is centred around data in PDB [2] [133], UniProt and other in-house data sources.

2.3 Conclusion

In the ambition to achieve computational modelling and inference of knowledge from data, the PPI data is not the only layer but the most important one for the multi-layer data driven models. Thus it is essential to link these multiple layers of heterogeneous data which then could be used as a set of predictive features for PPI prediction.

In this chapter widely used genomic features were presented. These feature have been employed in parallel to the known PPIs in order to exploit and infer the genomic similarity between interacting proteins and identifying features which distinguished them from non-interacting pairs. These features include – but are not limited to – sequence arrangements, sequence similarity, domain composition, co-evolution, co-expression, and GO similarity features. The core assumption of closely located genes (along the genome) should imply functional similarity, however such gene neighbouring was observed rather in prokaryotes but not yet in eukaryotes. Nonetheless, in those studies, only the linear, euclidean, distance between the two genes has been considered and defined as neighbourhood feature. It is only natural that in higher organisms the complexity of gene arrangements and protein functionality would scale up to higher degrees. Therefore, the simplest linear distance might not be the solution to the question of the relation between the genomic location of genes and the functionality of their proteins.

Next chapter of this thesis will investigate the possibility of an alternative representation of the genomic location which could relates the 2D linear chromosomal position of the genes to the 3D chromosomal folding. This alternative representation not only links 2D and 3D genomic location but also investigate the relation of the common partners of the proteins to their functionality and chromosomal folding.

[2] http://www.rcsb.org/pdb

Chapter 3

Prediction of Protein-Protein Interactions using 3D chromosomal locations

3.1 Introduction

Despite the extensive number of studies published on PPI network analysis our understanding of the human interactome remains incomplete. A large amount of relevant data is available as supporting evidence for interactions between two proteins. Such data includes the genomic distance between genes that code for two protein interaction candidates. It has been shown for prokaryotes that the intergenic distance between neighbouring genes is important for transcriptional activities [134]. Although the transcriptional operon for eukaryotes does not seem to exist, there is evidence which suggests that the genomic arrangement in eukaryotes is not random and that neighbouring genes are more likely to be co-expressed and functionally related [135]. Studies having focus on linkage between the chromosomal location of genes and their co-expression reported that genes which are involved in the same pathways are closely located on the genome which suggests co-regulation of co-functional genes [136]. In a similar way, multi-gene regulatory machinery has been observed in *Saccharomyces cervisiae* suggesting the existence of shared transcriptional regulatory elements for co-expressed genes which is equivalent to the concept of the operon in prokaryotes [137]. Also, a positive correlation has been reported between intergenic distances in comparison to distances between their products in PPI networks in yeast [138].

In several prediction base studies and tools gene neighbouring was used as a predictive feature. An example of such tools is Stringdb [139] which defines the gene-neighbouring

score as the inter-gene nucleotide counts. However, all the above mentioned studies only account for the linear distance which directly affect the co-evolution, co-expression and co-functionality of the neighbouring genes but the effects stretch out further than close proximity of the genes. Taking into account the linkage between close proximity and co-functionality as well as cluster of co-functional proteins in a network of PPIs lead to the question that whether there is an interconnection between intergenic distance of co-cluster proteins in a PPI network.

The work presented in this chapter investigates the relation between the genomic location of the neighbouring genes and their protein interactions. Thus, the distance between genes is translated to a new feature which is able to capture the nucleotide counts definition for protein partners located on the same chromosomes as well as strength of the linkage between partners which are located on two different chromosomes. Here the main assumption is the selective interaction between co-functional segments of the human genome. Since these segments might not be all located on the same chromosomes, the genomic location score which will be explained in the following section was introduced. The analysis has been extended to evaluate the role of such selective interactions for the prediction of PPIs.

3.2 Methodology

This section describes the methodology which has been used to classify a set of given interactions. The method involves the application of Naïve Bayes classifier on a set of biomedical features which have been extracted or calculated based on experimentally validated PPIs. In order to evaluate the predictive power of novel features with a state of the art approach, I applied the Naïve Bayes approach to multiple scenarios each being composed of different feature selections. Hence the results demonstrate the combinatorial effect of different features in the prediction of PPIs. Fig. 3.1 illustrates an overview of the model, of the features and their relation to the PPI prediction problem.

3.2.1 Prediction algorithm

The Naïve Bayes classifier has been applied as the prediction algorithm in order to predict the interactive or non-interactive class for a given pair of proteins. WEKA[1] package has been used in order to run the algorithm.

[1] `https://weka.wikispaces.com`

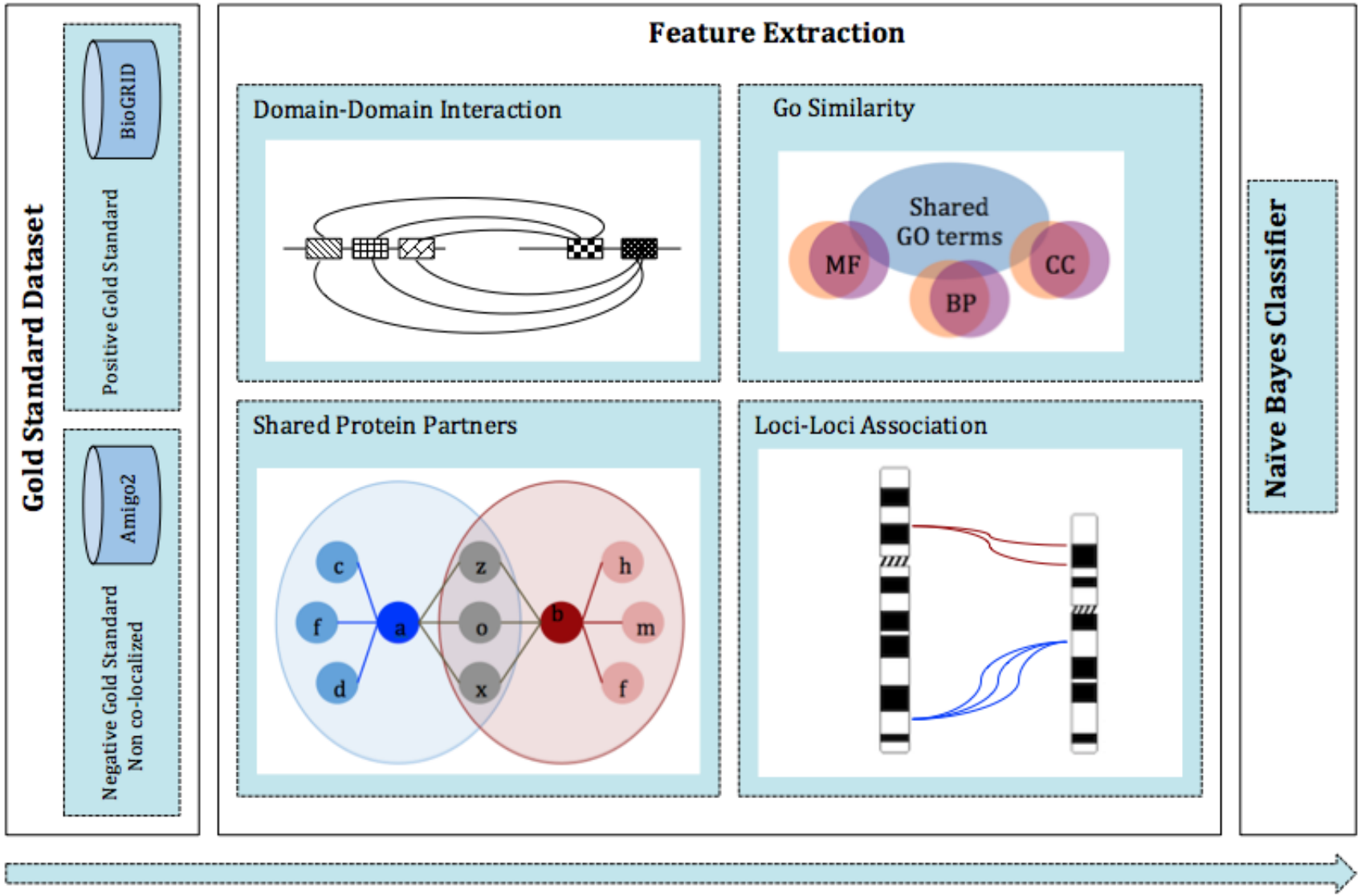


FIGURE 3.1: Landscape of the model.

A ten-fold cross-validation assessment was performed on the gold standard data which will be explained in more detail in the following section. In order to perform a ten-fold cross validation the gold standard dataset was randomly divided into 10 subsets and in each round of cross validation nine subsets were used as *training* sets and the one set which was left out was used as *test* dataset. The mathematical representation of the Naïve Bayes approach is as follow:

Given a set of independent variables or features:

$$X = \{x_1, ..., x_n\} \tag{3.1}$$

the goal is to learn the posterior probability of $P(X|C_i)$ where C_i belongs to the set of possible outcomes or classes $C = \{C_1, ..., C_n\}$.

The Bayes theorem is defined as follows:

$$P(C|x_i) = \frac{P(x_i|C)P(C)}{P(x_i)}, \tag{3.2}$$

where $P(C|x_i)$ is the probability that x_i belongs to class C. Assuming the independence between features or *the class conditional independence*:

$$P(x_i|C, x_i, ..., x_n) = P(x_i|C), \tag{3.3}$$

$$P(C|x_i, ..., x_n) = \frac{P(C)\prod_{i=1}^{n} P(x_i|C)}{P(x_1, ..., x_n)}, \tag{3.4}$$

The probabilistic model is:

$$P(C|x_i, ..., x_n) \propto P(C)\prod_{i=1}^{n} P(x_i|C), \tag{3.5}$$

the Naïve Bayes classifiers uses *maximum a posterior(MAP)* in order to estimate $P(C)$ and $P(x_i)|C$ then the relative frequency of class C in the training data set is derived as follow:

$$\hat{C} = argmax_k P(C)\prod_{i=1}^{n} P(x_i|C). \tag{3.6}$$

3.2.2 Gold standard dataset

The evaluation of a classification problem requires a gold standard data set which comprises training and test dataset. The gold standard dataset contains data points which

belong to any of the classes that are given in the prediction problem. For the PPI prediction problem, each data point belongs to either the interacting or the non-interacting class which represent the positive and the negative gold standard dataset, respectively.

Positive dataset: Experimentally validated interactions were retrieved from BioGRID (release 3.4.129) in order to form the core structure of the network. BioGRID has established strong partnerships for data exchange with providers of model organism databases as well as interaction databases, therefore it is the most comprehensive source for PPIs. The current prediction model is built on PPIs identified for humans. BioGRID contains both genetic and physical interactions, however in this work only physical interactions have been considered for the positive gold standard dataset. BioGRID provides additional information on the type of experimental method used for the verification of each interaction but the proposed model does not take the different experimental methods into consideration. In other words, all PPIs from BioGRID have been included in the positive gold standard dataset regardless of the experimental methods used, i.e. high- versus low-throughput identification methods. Moreover, the self-interacting pairs have been removed from the dataset. Eventually, the total number of 187,448 unique interactions have been collected as *positive gold standard dataset*, i.e. the *interaction dataset*.

Negative dataset: One of the major challenges in gold standard preparation in the classification of PPIs is the low coverage for *non-interacting* pairs in the primary data sources. The scientific literature mainly reports on interacting protein pairs thus there is no reference source for the provision of non-interacting proteins. Hence in PPI prediction studies, the negative gold standard datasets have been generated in an artificial way. The common approach to address the lack of a negative gold standard dataset is either: (1) a selection of interaction pairs from randomized proteins or (2) a selection of pairs proteins where evidence is given that the proteins do not co-localise. The former approach uses proteins from the positive gold standard dataset and produces random pairs, which then form the negative gold standard dataset. The later assumes two proteins interact only if they are residing in the same cellular compartment, therefore generates the non-interacting pairs based on the proteins' cellular locations, i.e. through the selection of protein pairs where the two proteins are known to locate in different compartments.

In this work, the negative data set was generated based on the non-localisation approach. The list of proteins which have been used in this work was retrieved from Amigo 2[2] based on their cellular compartment annotations. The generated list is composed of proteins which have been annotated to one of these three compartments: *nucleus*, *membrane* and *cytoplasm*.

[2] `http://amigo.geneontology.org/amigo`

The current modelling approach requires equal sizes for both, the positive and the negative gold standard dataset. Therefore a set of 187,448 unique protein pairs, equal to the size of the positive gold standard dataset, was generated as non-interacting or negative gold standard dataset.

3.2.3 Feature selection

Feature extraction has been carried out subsequently to the collection of the gold standard dataset and all PPI instances have been transformed to a feature vector. Table 5.1 lists every feature and indicates whether the feature has been identified by mapping the PPI instance to other data sources or has been calculated.

TABLE 3.1: Features and data sources.

Feature	Data source
Domain composition	UniProt, PFAM
Domain-Domain association score	Calculated
Go Similarity	Gene-Ontology
Ideograms	CellBase
Ideogram-Ideogram association score	Calculated
Neighbouring network	BioGRID

3.2.3.1 Domain-domain interactions

The domain compositional structure for each protein and for every PPI instance in the gold standard dataset has been extracted from the UniProt and PFAM databases. The domain-domain interaction features have been selected to give answers to the following two questions: (a) which two domains interact, and (b) what is the significance of the interaction between a pair of domains? Hence the domain-domain interaction feature has been chosen in a way that it embraces both the identity of domains and the significance of its interaction in a PPI instance.

In order to measure the significance of each domain-domain interaction, the association score has been calculated for each domain-domain interaction in the dataset. Here is a brief description of the domain-domain association score based on the scoring function proposed by Margalit et al. [140]. The main assumption is based on the fact that proteins differ in the compositional structure of the domains, i.e. the number and type of domains, and depending on the functionality of two proteins, they may share one or multiple domains with the partnering candidate. It is noteworthy that very few domain pairs for known protein pairs have been experimentally validated. However, it is possible

to infer domain-domain interactions from protein-protein interactions by identification of specific patterns. In other words, the frequency of two observed domain co-occurrences from the entire positive gold standard dataset has been used to determine such dominant domain interaction patterns.

Here we assume P_a and P_b are two proteins comprise of set of domains

$$D_a : \{D_1 ... D_k\} \tag{3.7}$$

and

$$D_b : \{D_1 ... D_m\} \tag{3.8}$$

The significance of the interaction between the two proteins based on their domain composition is calculated from the following formula:

$$SD_{a,b} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{n_{ij}}{n_i n_j} \tag{3.9}$$

where n_i and n_j are the number of times D_i and D_j were observed respectively and n_{ij} is the number of times D_i and D_j were observed together in a pair of interacting proteins.

In instances where one of the proteins has not been annotated with any PFAM domain the default value of *PF:000000* is assigned to the domain composition of the protein.

3.2.3.2 GO similarity

All the proteins in the gold datasets have been mapped to their gene-ontology (GO) annotations for *cellular components(CC)*, *molecular function(MF)* and *biological process(BP)*. In instances where one of the proteins is not annotated with any GO term the default value of *GO:000000* has been assigned for the GO category.

3.2.3.3 Neighbouring network

Proteins with a higher number of shared partners are more likely to be involved in the same cellular process and to expose a similar function. It is reasonable to assume that being involved in similar cellular processes requires a binding of proteins, their interaction or the formation of protein complexes. In order to test this premise, the neighbouring network of a protein has been analysed to add a neighbouring feature to the predictive features set. The neighbouring network of two proteins for each PPI instance for the gold standard data set has been defined as follow: If $N_i = \{P_1 ... P_m\}$

TABLE 3.2: Feature composition of different scenarios.

		Features Set					
		Domains	Domain Score	Ideograms	Ideogram Score	Common Partners	GOterms
Scenarios	A	*	*				
	B	*	*	*	*		
	C	*	*	*	*	*	
	D	*	*	*	*		*
	E	*	*	*	*	*	*
	F			*	*	*	
	G	*	*			*	

and $N_j = \{P_1...P_n\}$ are sets of proteins which directly interact with P_i and P_j. The neighbouring network set for each pair is the intersection between N_i and N_j:

$$(N_i \cap N_j) = \{P_1...P_k\} \tag{3.10}$$

3.2.3.4 Genomic location

In this model, the genomic location score for each protein pair from the gold standard dataset has been generated, not only based on the exact position of their coding genes but also on the ideogram on which the gene resides. Ideograms are patterns formed on stained and compact chromosomes and may cover several consecutive genes. In cases where the genes fall into two consecutive ideograms the one that contains the longer segment of the gene sequence was assigned to the genomic location of the gene. Similarly for the domain-domain interaction feature, the genomic location determines which of the two segments of the genome are interacting and how significant the interaction between these segments is. The location correlation has been measured based on the frequency of the ideogram co-occurrence for pairs in the interaction data set:

$$SI_{i,j} = \frac{n_{i,j}}{n_i n_j} \tag{3.11}$$

where i and j are the pairs of ideograms, $n_{i,j}$ is the frequency of the pair's occurrence, n_i is the frequency of i and textitn_j is the frequency of j in the entire dataset. SI is the confidence ratio score for the interaction between i and j.

3.2.4 Gene ontology enrichment analysis

GO enrichment was performed using the R package *topGO* with Fisher's exact statistical test. The cut-off p-value has been set to 0.05.

3.3 Results

Seven different scenarios were used for the problem of classification of PPIs. These scenarios include the different combinations of feature sets which have been explained in the previous section. In order to compare the relevance of the genomic location score feature and the network of common partners in the classification problem, the domain composition and domain association scores have been jointly considered as the baseline features. These two features have been successfully analysed in other related studies where the application of machine learning approaches have been explored in similar ways. Therefore, this approach is a reliable choice for the baseline scenario in order to compare the predictive power of other features. The results from the PPI prediction for the different scenarios is given in table 3.3.

The highest accuracy of *94.81%* has been achieved in *scenario F* where the combination of features included the genomic location and the network of shared partners in the predictive features set. This result indicates the correlation and thus the potential relation between the genomic location of a protein and its interaction partners. From the network analysis perspective, it has been shown that functionally related proteins form a clique in the network of interactions [141]. In fact, one way of functionally annotating proteins having unknown functionality is to annotate them based on the known functionality of their interaction partners. In other words, these results suggest that functionally related proteins are positioned in genomic locations with relevance to the same functionally. However, functionally related loci might not be two consecutive loci or even on the same chromosomes. Thus these results indicate the preferential interactions between different genomic loci in favour of a shared functionality.

On the other hand, studies have shown that genes with similar co-expression patterns tend to be involved in similar functions. Similarly, genes residing in close proximity show higher co-expression correlations. These evidences in combination with the result of *scenario F* insinuate the inter- and intra-chromosomal preferential interactions among segments of the genome.

As a consequence, the GO enrichment analysis has been performed for each chromosome. The heatmap which is shown in Fig. 3.2 illustrates the GO term annotations in relation

to the chromosomal location. As can be seen, the tiling pattern characterises a set of distinct cellular processes for each chromosome with a small overlap between chromosomes. These observations point towards and stress the relation between chromosomal folding and functionally related loci. The 3D contact map for the genome proposes chromosomal territory with small chromosomes folding in the inner cycle of the nucleus and the bigger chromosomes loosen around them and forming a globular structure.

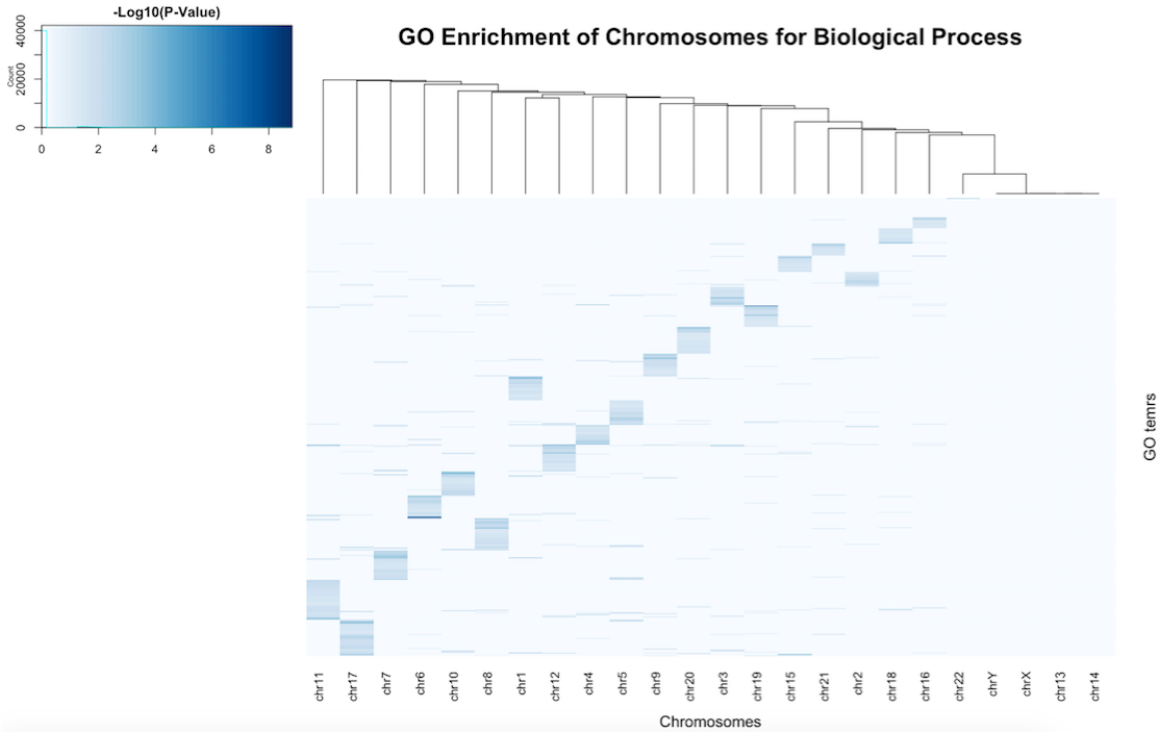


FIGURE 3.2: GO enrichment of chromosomes for biological process (P-value ≤ 0.05).

The chromosomal territory and the distinguished set of cellular processes for each chromosome as shown in Fig. 3.2 in combination with the prediction result of *scenario F* suggest (a) inter-chromosomal preferential interaction and (b) intra-chromosomal selective interactions. However, further studies are required in order to investigate the association of the 3D contact map with the human genome and human interactome.

The second best performance,*93.51%* results from *scenario G* which integrates the domain composition, domain association score and the network of shared partners. As mentioned above proteins with similar functionality form functional cliques in the interaction network. Therefore it is expected that proteins which are composed of functionally related motifs are functionally related to each other in the network. As a result, the neighbouring network and domain-domain interaction features are complementary features and if used together will improve the performance of the model. However, this does not suggest that these two features are similar and their combinatorial application is not against the fundamentals of feature selection.

Scenario C and *Scenario E* performed in a similar way with slight advantages for *Scenario E* where the GO similarity of pairs has been taken into accounts. It can be assumed that the reason for this minor improvement is the incompleteness of the GO annotations for the protein datasets which imposes a significant volume of missing features onto the model. One way of resolving this issue would be to include only protein pairs with all available features however this would reduce the number of interactions in the gold standard dataset to a large degree. Instead – as mentioned in the methods section – a default value was introduced for the missing GO terms in order to be able to keep all the interactions in the gold standard dataset and use their other features for prediction of PPIs.

However comparing *Scenario C* with *Scenario F* and *Scenario G*, one would expect a better performance of the model since *Scenario C* embodies the other two scenarios. *Scenario E* and *Scenario F* are based on the neighbouring network parameter in combination with domain composition and genomic location. Comparing these two to *Scenario B*, which is composed of the union of genomic loci and domain composition, lead to the reduction of the model's accuracy. Taking these four scenarios and apparent observations all into account leads to the conclusion that the association of genomic loci and domain composition is less relevant. This outcome in itself is not surprising, however, the interrelation of both of these features in combination with neighbouring network feature leads to concerns. Further functional studies are needed in order to explain this observation.

Lower performance accuracy has been obtained when the shared partners have been removed from the predictive feature set in *scenario D*. Similar to *scenario E*, the amount of missing GO annotations could be the partial reason for the decrease in accuracy. Also, based on other scenarios which were discussed above, there is a connection between the shared partners of proteins and their genomic location as well as their functional building blocks, their domains. Therefore, omitting the shared partners from the predictive feature set caused the decline in performance of the approach. This result suggests that the neighbouring network feature forms the key feature from the whole set to achieve higher prediction accuracies in combination with other features.

3.4 Discussion

This chapter presented an integrated feature-base approach for the prediction of PPIs. The use of a NB machine learning approach based on genomic properties of the proteins enabled the prediction of PPIs. Based on the results which were presented in the previous section the interacting network similarity of two proteins has the highest impact on the

TABLE 3.3: Average model performance on different feature scenarios.

	Accuracy	Precision	Recall	F-measure	AUC
Scenario A	91.66	92.82	91.66	91.6	0.99
Scenario B	88.84	90.69	88.84	88.71	0.98
Scenario C	91.01	92.24	91.01	90.94	0.99
Scenario D	89.40	91.07	89.40	89.29	0.98
Scenario E	91.36	92.47	91.36	91.30	0.99
Scenario F	94.81	95.14	94.81	94.8	0.99
Scenario G	93.51	94.23	93.51	93.48	0.99

classification of PPIs. In other words, the higher the number of shared network partners for a pair of proteins the higher is the likelihood of the two proteins to interact. This is in accord with the studies which used the protein partners as a mean for the functional annotation of proteins.

Also, these results demonstrate the association between the genomic location of the proteins and their partners and thus promotes the idea of regulatory mechanisms for eukaryotes that resemble the operon like machinery in prokaryotes. However, the complex structure of the eukaryotic genome may significantly differ from the linear approach given for the operon concept. In more detail, the proximal and distal distance in the eukaryotic genome is determined through a 3D spatial representation instead of 2D linear one. Although the existence and relevance of the transcriptional regulation in eukaryotes is widely acknowledged, a comprehensive understanding of the regulation in relation to the protein's genomic location is yet not available.

On the ground of the results presented above, it can be concluded that the genomic location of proteins form the core for future prediction methods for PPIs. Thus, further research on the 3D genomic location of interacting proteins with regard to their position in the nucleus might introduce further relevant features for the PPI prediction problem.

Recent studies published [142] the 3D contact map of the genome in 8 different cell types. The 3D contact map illustrates the chromosomal folding in the nucleus. Chromosomal folding brings the distal segments of the genome to the close proximity. Each chromosome folds in its own territory, therefore the physical contact within chromosomes is expected to be more extensive than the physical contact between chromosomes. The mapping of known PPIs to their genomic location together with the available transcriptional network will enable us to infer the association between protein partners, their transcriptional mechanism and their nucleus position.

Last but not least, in this study it has been demonstrated that for each chromosome exists an enrichment of annotations with regards to cellular processes which is distinctively

different from the other chromosomes. Some overlaps in the annotation exist between chromosomes for selected enriched cellular processes.

In combination with the 3D folding data of the chromosomes, future research work will be able to explain the formation of functional cliques in the PPI networks. 3D contact map of human genome investigates the physical contacts between segments of the DNA on a folded chromosome. However, these contact maps generate high volume of data which makes it difficult to use the contact scores as a predictive feature with traditional approaches. Thus, the choice of predictive models and their development are important. Future work will focus on how to successfully apply these 3D contact map information in combination with the regulatory information as predictive features.

The majority of the content of the following chapter is re-printed from the paper **"LinkedPPI: enabling intuitive, integrative protein-protein interaction discovery.** *Laleh Kazemzadeh, Maulik R. Kamdar, Oya Deniz Beyan, Stefan Decker, Frank Barry."*

Chapter 4

Semantic Integration

4.1 Introduction

The study of biological networks forms an integral core of biomedical research related to human diseases and drug development. The ultimate goal of such studies is to understand the connections between different genes and proteins, how the cell signals propagate across these networks and regulate the cell's functionality. Hence, understanding the PPI networks underlying such cellular mechanisms is important to identify dysfunctions linked to specific proteins leading to the discovery of novel drug candidates and treatments for diseases. Studying PPI networks facilitates the understanding of the interconnectedness of different cellular mechanisms and pathways. In contrast to past beliefs, biological pathways are not independent of each other but their meshing is harmonious, which makes them components of a bigger maze. Thus, it is important to investigate the dynamics of the cell system as a whole in contrast to a selection composed of single modules.

The major challenge in developing a thorough understanding of these cellular mechanisms and pathways is to generate a complete representation of the PPI network for each biological mechanism. Considering the high order of interactions amongst proteins in the human genome, additional assistance is needed in order to narrow down the number of hypothetical interactions. Computational models mitigate the identification of the sequential, structural and physio-chemical properties of known interacting protein pairs and highlight the underlying patterns. As a result, wet-lab validation of the hypothesis derived from the computationally extracted links and protein partners becomes realistic, achievable and suitable for efficient identification of novel interation pairs.

In the era of *'-omics'* technologies, data integration approaches form the right answer for the inundating amounts of biomedical data. Data Integration approaches open the

boundaries between multiple research disciplines by generating the connectivity in multi-dinsciplinary research domains. The employment of integrative approaches enables researchers to benefit from the combination of the multitude of *auxiliary data* coupled with the domain peculiar data, in forming the hypothesis and venturing *in silico* models. However, integrative modelling appears to be a demanding task. In the case of PPI analysis an integrative model profits from pertaining data (e.g: gene-expression). However, generally biomedical data has been published to address specific, albeit very different research problems. Therefore the data representation, data model and formats may vary from one data source to the other. Challenges stemming from the heterogeneity of the data emphasise the need for a framework which can bridge between these biologically different concepts in order to highlight and extract the ubiquitous patterns, inconspicuous in the *bigger picture*. Nevertheless, these obstacles originate from the interdisciplinary nature of integrative modelling. Here the major plights are the connectedness between similar concepts in a single domain and conceptual unification of terminologies between domains. These hurdles lead to gaps in knowledge and consequently impede receiving the utmost benefits offered by the deluge of biomedical data.

The main obstacle in integrative approaches is the *'language gap'* [143]. The language gap refers to the lack of understanding of terminologies of different or disjoint research domains. Therefore, integration of the multitude of data appears to be problematic. Moreover, the traditional integrative approaches (e.g: local data warehousing) fails to meet the comprehensive sharing principles. Thus a new generation of integrative technologies based on Semantic Web technologies and Linked Data principles can potentially address these issues.

Semantic Web and Linked Data are closely related with the constructs of *'ontologies'*. An ontology is a special form of data representation with a clear focus to formalise the conceptualisation of knowledge from observations in the real world. A priori, an ontology is a set of terms (and their definitions) which denote concepts and their associations in a specific scientific or non-scientific domain, and in addition provides formal constructs to denote the relations between concepts and even between different ontologies. The inherent interoperability of ontologies reduces the gaps between multiple (scientific) domains which forms the key gateway for integrative approaches. However, every single ontology is constructed to represent a specific domain, therefore, reproduces the domain's complexity in its representation. While the granularity and size, i.e. the level of detail, of an ontology is considered to be an advantage, often integrative approaches do not require the complexity and large volume of terminologies encompassed in each ontology. Thus a less complicated and more domain specific set of terms which are capable of presenting and modelling the domain specific data is often more practical and advantageous.

The main motivation of the work presented in this chapter is to enable researchers through a framework of integrated data sources to retrieve the answers to their research questions from these disparate data sources. A researcher interested in the list of protein domains in a specific protein can look them up at the UniProt website[1] which provides rich information on proteins. Genomic locations of protein-coding genes are publicly available from several web services such as CellBase [144]. However requests like, *'List of all the proteins which contain the exact or partial set of protein domains?'* or *'What is the relation of a set of interacting proteins and the genomic location of their underlying genes?'* cannot be answered through these websites.

4.2 Methods

The final goal of this research is the identification and extraction of potential PPI networks from publicly available but distributed data sources. The core structure of the PPI network consists of proteins and their experimentally validated interactions. Fig. 4.1 depicts an overview of the LinkedPPI architecture. The following subsections will describe the data selection, RDFization and integration methodologies used.

4.2.1 Selection of relevant data sources

LinkedPPI aims to extract relevant protein pairs based on the set of protein features. The foundation of LinkedPPI is based on collecting all experimentally validated interactions. The protein features in LinkedPPI cover the extended data from complementary data sources. This section explains the detail description of data sources which have been used for LinkedPPI.

4.2.1.1 Validated interactions as a dataset

Experimentally validated interactions were retrieved from Biological General Repository for Interaction Datasets (BioGRID), one of the most comprehensive PPI databases [145]. Only physical interactions were included in this work, regardless of their classifications as low or high throughput. The non-redundant and self-interaction records were omitted from the dataset.

[1] `http://www.uniprot.org`

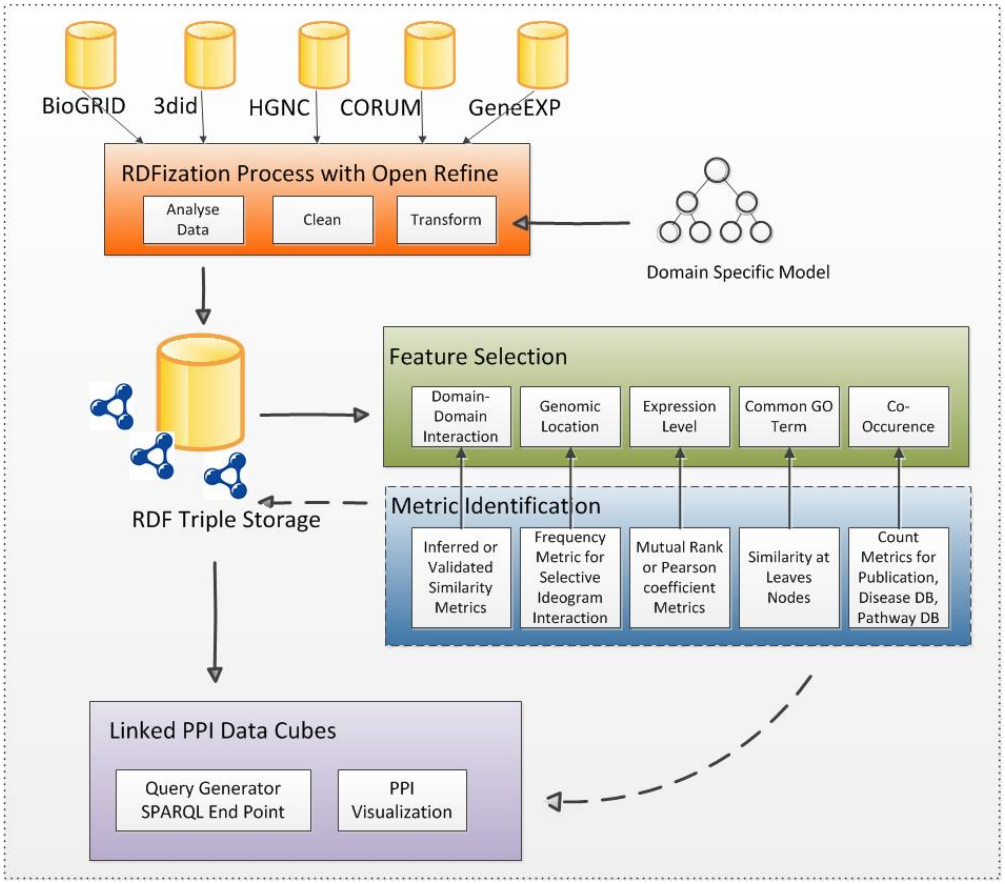


FIGURE 4.1: LinkedPPI Architecture

TABLE 4.1: Integrated data sources.

Data type	Data source
Validated Interactions	BioGRID
Protein complexes	CORUM
Gene expression	COXPRESdb
Genomic Location	CellBase
	Calculated
Domain composition	UniProt
	PFAM
Domain-Domain interaction	3did
	Calculated
GO Annotations	Gene-Ontology
Gene co-occurrence	calculated

4.2.1.2 Protein complexes

In cellular processes proteins act as a complex, in contrast to a binary interaction between a pair of individual proteins [146], and thus act during the same time and within the same cellular compartment. These proteins are tightly interacting and play key roles in PPI networks. Elucidation of the dynamics of PPI networks and functionality of individual

proteins can lead into the identification of essential protein complexes since different subunits contribute toward the same cellular function. In this work, Comprehensive Resource of Mammalian protein complexes (CORUM) [147] was used as the primary source for protein complex data.

4.2.1.3 Gene expression

Understanding the correlation between gene-expression networks and protein interaction networks is an ongoing challenge in PPI studies. Proteins encoded from co-expressed genes are more likely to interact with each other [148]. The COXPRESdb data [149] which publishes any recent gene expression microarray results has been used as the source for co-expressed genes. COEXPRESdb provides correlation data for expression results for several model organisms however only human data has been used in this work.

4.2.1.4 Genomic locations

Neighbouring genes show similar expression patterns and are often involved in similar biological functions [150] which suggests that they might share same activation and translation mechanisms. These interactions may not be limited to the adjacent genes but can be long-range interactions to fulfil the cellular functionality [151]. Such evidences introduce another layer of data for the genomic locations of the protein-coding genes in our framework. Here the 'genomic location' is not defined as the exact start and stop position of genes on a chromosome, but as the Ideogram band in which the genes reside. Ideograms are schematic representations which depict fixed staining patterns on a tightly coiled chromosome in karyotype experiments. A karyotype describes the number of chromosomes, their shape and length and banding patterns of chromosomes in the nucleus.

The Ideogram data was downloaded from the mapping and sequencing tracks in the Human Genome Assembly (GRCh37/hg19, Feb 2009) at the UCSC Genome Browser[2]. The start/stop coordinates of the genes were retrieved from CellBase [144] and used to determine the genes within each ideogram.

4.2.1.5 Protein domains

The protein's functionality and its structure is defined by its specification of domains. The sharing of domains between proteins indicates to interaction of proteins with similar partners. However, the identification of the correct interactions between domains

[2] `http://genome.ucsc.edu/`

through the experimental validation for all possible protein pairs is an insurmountable task. Therefore, mainly the integration of knowledge bases for protein domains advanced the research on the identification of more complete PPI networks. Similarly, the integration work contributes toward the identification of novel domain-domain interactions without requiring further experiments (e.g., X-ray crystallography). The Database of three-dimensional interacting domains (3did) was used for the LinkedPPI platform [152], since it provides high resolution three-dimensional data for structural interactions of domains. The structural composition of domains for each protein has been extracted from the UniProt knowledge base.

4.2.1.6 Gene co-occurrence

The large portion of the published research work in the biomedical research domain has been accumulated typically around a specific set of genes in the context of a pathway or in the association to a predefined disease. Thus, studying the co-occurrence of genes across the scientific literature [153] forms another dimension to the identification of PPIs, similar to pathway and disease association studies. Previously Kamdar et al. generated co-occurrence scores as a weighted combination of the total number of diseases, pathways or publications in which any two genes occur simultaneously [154]. This co-occurrence score was retrieved for each pair of proteins in the core network of LinkedPPI which was constructed from the BioGRID data.

4.2.2 Entity mappings

To determine which gene is responsible for the encoding of a given protein, i.e. to determine the mapping between the Entrez-Gene ID and the UniProt ID, the ID mapping table[3] provided by UniProt has been exploited. One of the major advantages of using this approach was that the mapping was linked to the relevant Gene Ontology (GO) [25] terms and the Entrez-Gene ID, thus providing additional information regarding the localisation and functions of the specified genes. The HGNC (HUGO Gene Nomenclature Committee) has been used to map common genes referenced by different identifiers (Entrez-Gene, Ensembl and UniProt) [155].

[3] `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/README`

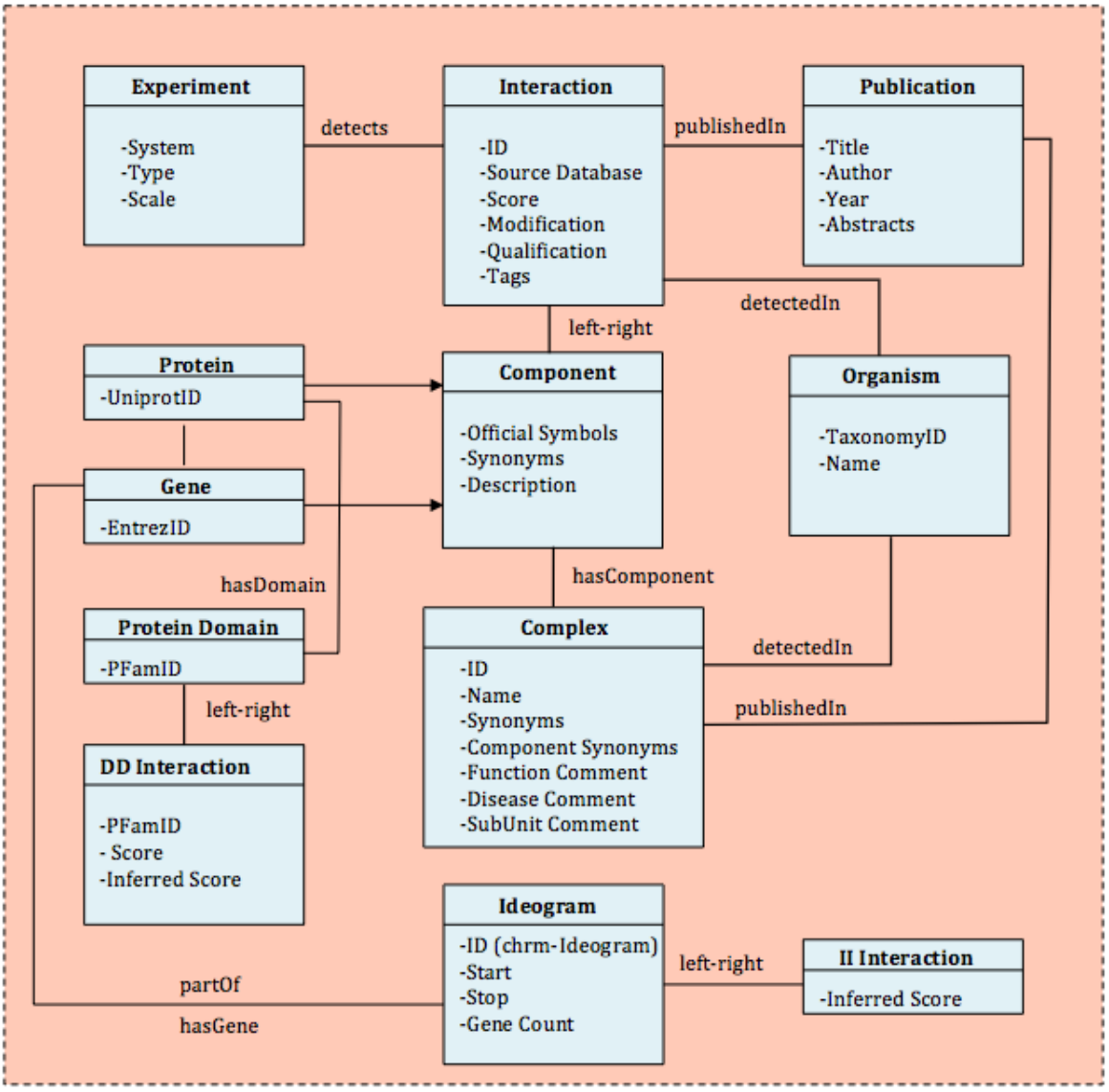


FIGURE 4.2: Class Diagram of LinkedPPI Domain-specific Model

4.2.3 The LinkedPPI domain specific model

One of the crucial challenges in integrative bioinformatics is the heterogeneous nature of biological data sources. Even though several attempts have been made in the standardization of the data through controlled vocabularies and guidelines, various hurdles still need to be surpassed. The proteomic standards initiative-molecular interaction (PSI-MI) is widely accepted by the community for the modelling of biological networks [31]. Even though some of the data sources are represented using the PSI-MI format there is no decipherable interconnectedness between them. To introduce the desired interconnectedness or 'bridges' between these data sources, the principles of Linked Data technologies have been applied for data integration and data interoperability. A simple and concise domain model has been used for the modelling of the PPIs retrieved from

BioGRID, the complexes from CORUM, the protein domains and genomic location. A domain-specific model is beneficial in comparison to an extensive well-constructed ontology due to the absence of non-domain-specific concepts (`Thing`, `Continuant`, etc.) and has proven to be much smaller and self-contained, which in return is more efficient to address a specific problem. Being native to a particular domain (e.g., *Protein-protein interactions*), it serves as an intermediate layer between the user and the underlying data, and enables intuitive knowledge exploration and discovery [156].

The following section describes the model which has been used to transform the data into a Linked Data format.

4.2.3.1 Description of concepts

The domain specific model which is proposed in this chapter comprises 12 concepts, which are crucial for the representation and exploration of proteins in PPI networks. Fig. 4.2 gives and overview on all concepts and their connections. Each component is formally described in the following paragraphs using the relations *is a*, *part of*, *interact* and others.

Component: The core concept in this model is a *Component*. A *Component* can either be a *Gene* or a *Protein*. A *Component* can be part of a *Complex* or can interact with another *Component* through an *Interaction*.

Gene: A *Gene* is identified by its Entrez identifier. A *Protein* has a *Gene* which is translated into the *Protein*. Also, a *Gene* is part of an *Idoegram*. The *Gene* is a *Component*.

Ideogram: A *Gene* is contained within an *Ideogram* therefore the *Ideogram* has a *Gene*. The *Ideogram* is identified by the combination of the chromosome and the ideogram band. The attributes of the *Ideogram* include the start and stop position on the chromosome as well as a number of genes which reside on the *Ideogram*.

Protein: A *Protein* is identified by its UniProt identifier and is a *Component*.

Protein Domain: The *Protein Domain* is associated with the *Protein* and is represented using the Pfam identifier of each domain.

Organism: The *Organism* denotes the species in which both the *Interaction* and the *Complex* have been determined. The taxonomic identifier of each organism and its label is attributed to this concept.

Experiment: The *Experiment* concept embodies the attributes related to the experimental *System* which has been used to detect the interaction (e.g. Y2H, AP-MS), the *scale*

of the experiment (high or low-throughput), and what *Type* of interactions it represents, i.e. physical versus genetic interaction.

Publication: The *Publication* specifies the resource described in the PubMed repository, where the experiment has been published, and references the *Title*, *Author*, *Year* and *Abstract* of the publication.

DDinteraction: The concept *DDinteraction* covers the association scores between two domains retrieved from 3did for PPIs in BioGRID. This is a pre-calculated score which has been included in the LinkedPPI dataset.

IIinteraction: The concept *IIinteraction* represents inferred interactions between two ideograms from experimentally validated PPIs. In the same way as *DDinteraction*, this concept covers the previously calculated association score which has then been integrated in the LinkedPPI data set.

4.2.3.2 Data transformation

In order to transfer the raw data to the RDF format, Open Refine has been used. Open Refine provides a workbench to clean and transform data and eventually export it in the required format. The RDF extension [157] of Open Refine was used to model and convert the tab-delimited files acquired from CORUM, BioGRID, 3did and CellBase into RDF graphs. The RDFized data is stored in a local Virtuoso triple store[4]. Data from COXPRESdb has already been published in RDF format therefore it was re-used considering its original data model and URIs. For the other data sources, i.e. for Entrez-Gene, UniProt and PubMed, the URIs for genes, proteins and publications have been reused in the same way.

4.3 Results

4.3.1 Statistics on RDFized data

After RDFization, the BioGRID data source formed the origin of about 11 million triples (11,357,231), which covered 634,996 distinct interactions between 14,135 Human proteins. The data source also links out to 38,952 unique PubMed publications referencing these PPIs. The CORUM data source consists of 156,364 triples from 2,867 distinct complexes. The 3did data source consists of 320,690 triples from 6,818 distinct protein domains and 61,582 validated and inferred domain-domain interactions. A total number

[4] `http://vmlifescience01.insight-centre.org:8890/conductor/main_tabs.vspx`

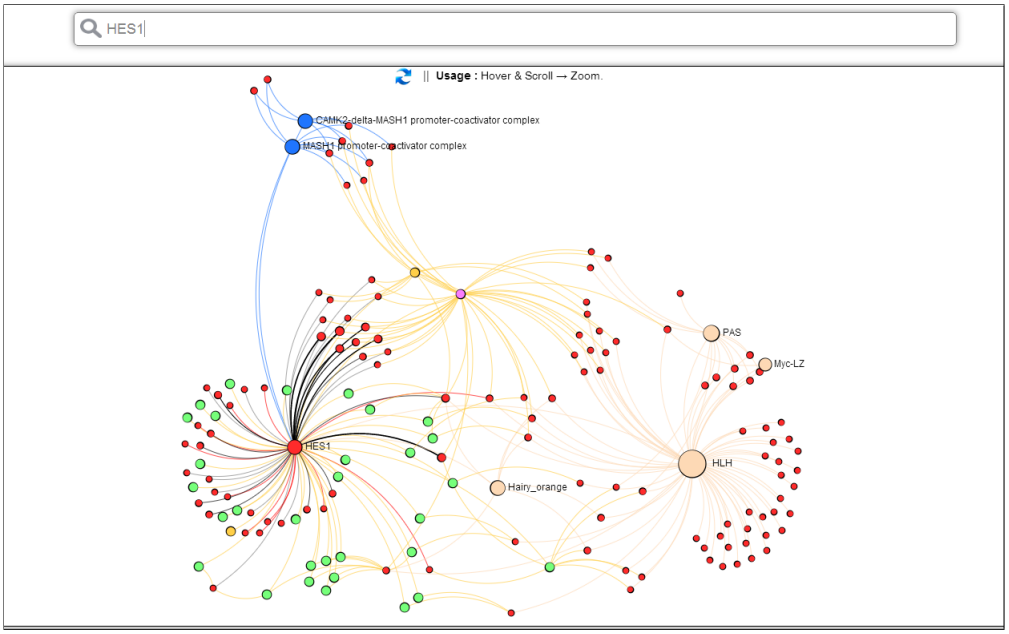


FIGURE 4.3: Searching HES1 protein using the PPI visualisation dashboard

of 13,493 interactions between 405 ideograms have been inferred which are referenced through 80,092 triples. 60,676 mappings were instantiated between the genes and 7,750 extracted GO child leaf terms. As such, relevant information could be retrieved from the SPARQL endpoint through the formulation of suitable queries.

4.3.2 Visualization in LinkedPPI

The preparation of SPARQL queries requires a domain knowledge and expertise, by contrast a biological or medical user needs an intuitive and interactive visualisation tool, which aggregates the information from the multiple data sources and represents the result in a combined view. LinkedPPI provides a PPI Visualisation Dashboard[5] based on Real-time Visual Explorer and Aggregator of Linked Data (ReVeaLD) [156] to accommodate the requirements for the search and visual exploration of the LinkedPPI networks.

Based on the official symbol of the desired protein a list of possible alternatives will be retrieved from the indexed entities. On selection, the entity's URI is passed as a parameter through a set of pre-formulated SPARQL SELECT queries[6] targeting the various data sources. As shown in Fig. 4.3, the PPI network associated with the searched protein (e.g., HES1 *entrezgene:3280*) is rendered in a force-directed layout. The list of entities retrieved from the data sources are represented as circular nodes with the size

[5] `http://vmlifescience01.insight-centre.org/linkedppi/`
[6] `https://gist.github.com/maulikkamdar/a47fbecddecc6ba4b373`

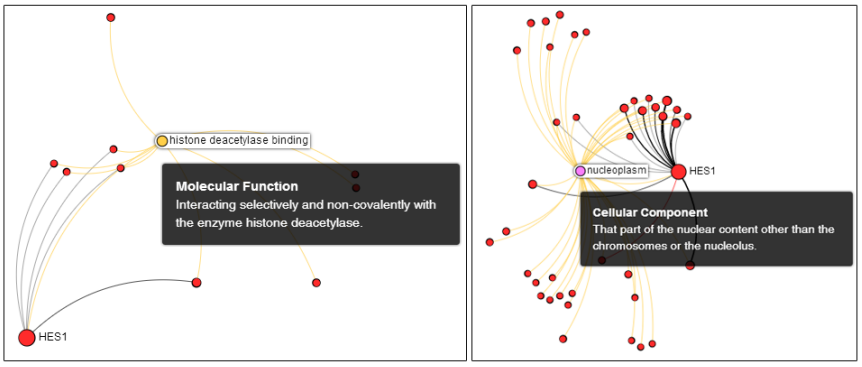


FIGURE 4.4: Subgraph of HES1 PPI network based on GO terms

of each node being proportional to the number of associated nodes. The nodes are rendered using different colors for the sake of visual differentiation – *Red* for *Components* (*Proteins* of BioGRID or *Genes* of COXPRESdb), *Blue* for CORUM *Complexes*, *Light Brown* for *Protein Domains* of PFAM and domain-domain interactions of 3did. The three categories of GO Child Nodes – *Biological Processes*, *Molecular Functions* and *Cellular Components* – are displayed using *Green*, *Yellow* and *Purple* colours, respectively. The interactions between different proteins are represented as edges – the colour of the edges is directly dependent on whether the associations have been retrieved from BioGRID, COXPRESdb or Co-occurrence Data (*Black*, *Red* and *Purple* respectively). The thickness of *Black* edges depends on the total number of publications which have experimentally validated the underlying interactions. The thickness of the *Red* and *Purple* edges depends on the Pearson Correlation Coefficient (PCC) for Gene expression and Co-occurrence scores respectively. The *Protein* nodes, which are present in the same complex, possess interacting domains or have underlying coding genes associated to the same GO terms, are not connected directly to each other by edges. They are all connected using similar coloured edges to the respective node (complex, domain or GO term), however there may be instances with experimental interactions or co-expression between the connected entities. The resulting network is hence densely clustered, rather than a simplistic radial layout of nodes. The listing 4.1 represents the query which fetches the PPI data, the listing 4.2 is the correspondence query to the list of co-expressed genes, the listing 4.3 retrieves the complexes in which the specified protein is involved, the listing 4.4 demonstrates the relevant query to the domain composition of the selected protein and the listing 4.5 extracts the Gene-Ontology annotations for the specified protein.

Hovering over a node highlights the subgraph of the network which only displays the first-level connected nodes and their relations (Fig. 4.4), hence allowing any domain user to intuitively deduce answers to simple questions like: *'Which protein-encoding genes in the network share the same molecular function and have experimental co-expression?'* An

information box is also displayed beside the hovered node to show additional information such as GO term descriptions, Pfam or PubMed IDs, and PCC scores.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT * WHERE {{
                ?biogridInt ppi:left entrezgene:3280; ppi:right ?geneTos.
        } UNION {
                ?biogridInt ppi:right entrezgene:3280; ppi:left ?geneTos.
        }
        ?geneTos ppi:officialSymbol ?geneSym .
        OPTIONAL {?biogridInt ppi:published_in ?pub}.
        OPTIONAL {?biogridInt ppi:qualification ?qual}.
        OPTIONAL {?biogridInt ppi:confidence_score ?score}
}
```

Listing 4.1: Query retrieving the validated PPIs.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX ns1: <http://coxpresdb.jp/rdf/def/0.1/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT DISTINCT ?it ?name ?pcc WHERE {{
                ?coexpressInt ns1:gene_id_1 entrezgene:3280;
                ns1:gene_id_2 ?it; ns1:pcc ?pcc
        } UNION {
                ?coexpressInt ns1:gene_id_2 entrezgene:3280;
                ns1:gene_id_1 ?it; ns1:pcc ?pcc
        }
        ?it ppi:officialSymbol ?name
} ORDER BY desc(?pcc)
```

Listing 4.2: Query retrieving co-expressed genes.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT * WHERE {
        ?corumInt ppi:componentsSynonyms entrezgene:3280;
        ppi:componentsSynonyms ?k; ppi:name ?name.
        ?k ppi:officialSymbol ?geneSym .
        OPTIONAL {?corumInt ppi:published_In ?pub}.
        OPTIONAL {?corumInt ppi:functionalComment ?comment} .
        FILTER (?k != entrezgene:3280)
}
```

Listing 4.3: Query retrieving complexes.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT DISTINCT ?name ?domain ?interDomain ?intername ?interGene  ?geneName WHERE {
        ?uniprot ppi:entrezGene entrezgene:3280  ; ppi:hasDomain ?domain.
```

```
        ?domain ppi:officialSymbol ?name .
        {
          ?inter ppi:left ?domain; ppi:right ?interDomain
        } UNION {
                ?inter ppi:right ?domain; ppi:left ?interDomain
        }
        ?interGene ppi:hasDomain ?interDomain .
        ?interDomain ppi:officialSymbol ?internname .
        ?interGene ppi:entrezGene ?gene.
        ?gene ppi:officialSymbol ?geneName.
        ?inter ppi:score ?score .
}
```

Listing 4.4: Query retrieving domain interactions.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT * WHERE {
        entrezgene:3280 ppi:GO ?go.
        ?geneComp ppi:GO ?go.
        ?go ppi:name ?name; ppi:namespace ?namespace; ppi:description ?description .
        FILTER (?geneComp != entrezgene:3280)
}
```

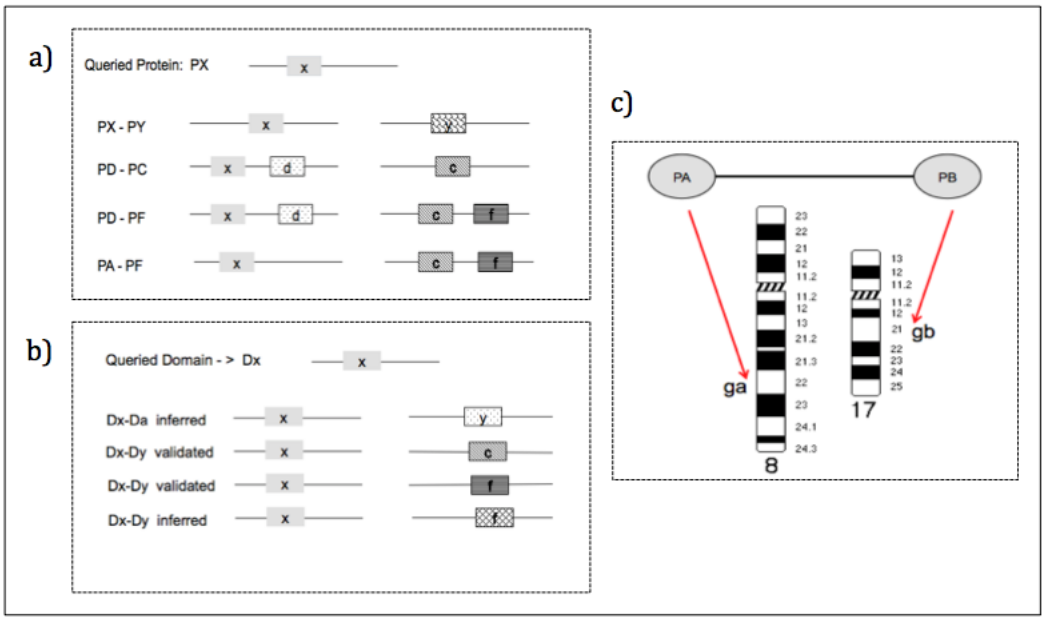
Listing 4.5: Query retrieving GO terms.

4.3.3 Use Cases

The following subsection describes three different scenarios (depicted in Fig. 4.5) where the LinkedPPI framework would provide support to facilitate the extraction of implicit information which can be used as predictors of novel protein-protein interactions. The relevant SPARQL queries are documented for each of the presented use cases.

4.3.3.1 Use Case 1: PPI candidates based on domain-domain interactions

This scenario aims to extract possible PPIs based on the known domain-domain interactions. For the sake of simplicity this use case focuses on proteins with single domains only. Considering a selected *protein* for which the sequence specification and domain composition are known, an interesting question is, *What is the list of potential protein partners for this* protein *based on the domain composition?* LinkedPPI can retrieve the list of protein pairs in which at least one of the proteins contains the same protein domain as the protein under question. Possible outcomes are as follow: (a) the protein under investigation itself appears in the result set which forms the list of its experimentally validated protein partners. However, this is already achievable from the BioGRID web



a) Use Case 1: *Px* denotes the protein under scrutiny which contains the domain *x*. **b) Use Case 2:** *Dx* denotes the domain of interest independent of any protein that would contain the domain. The returning result is a list of binary interactions between domain pairs, where each interaction is labeld either as *validated* (data retrieved from 3did) or *inferred* (data deduced from PPI data). **c) Use Case 3:** The genomic location of *PA* and *PB*, two interacting proteins, are mapped to their ideogrammatic location on two different chromosomes.

FIGURE 4.5: Illustration of the three use cases.

service directly. (b) List of all proteins with one single domain. In this case with a naive and straightforward conclusion, the researcher may accept the list. However in most cases further use of GO enrichment or advanced statistical analysis will offer a more concise list but these analyses are beyond the scope of the current LinkedPPI. (c) The query results to a set of proteins consisting of several domains which requires further statistical or domain expert knowledge refinement in order to identify which domains of the two proteins are involved in specific interactions.

Despite the need of further investigations in such cases, the shortlisted hypothetical interaction partners are expected to be brief to save a tremendous amount of time and laboratory work. The SPARQL query focuses on the example of the HES1 (*entrezgene: 3280*) protein. It obtains the list of domains present in HES1 which includes Hairy_Orange (*pfam:PF07527*) and HLH(*pfam:PF00010*), and the list of proteins (e.g. HEY2) which share either of these domains, or have domain-domain interactions. Then it retrieves validated PPIs in which the protein participates (e.g. HEY2-SIRT1). Also it acquires the PubMed publication documenting each PPI. The relevant query is shown in listing 4.6.

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT DISTINCT  ?requestedDomainName ?interactingDomainName
?interactingProteinName
   ?validatedProtein ?validatedProteinName ?publication WHERE {

        ?uniprot ppi:entrezGene entrezgene:3280; ppi:hasDomain ?requestedDomain.
        ?requestedDomain ppi:officialSymbol ?requestedDomainName .
        {
                ?interaction ppi:left ?requestedDomain; ppi:right
                ?interactingDomain
        } UNION {
                ?interaction ppi:right ?requestedDomain; ppi:left
                ?interactingDomain
        }

        ?interactingProtein ppi:hasDomain ?interactingDomain;
        ppi:entrezGene ?gene.
        ?gene ppi:officialSymbol ?interactingProteinName .
        ?interactingDomain ppi:officialSymbol ?interactingDomainName .
        ?interaction ppi:score ?score .
{
                ?biogridInteraction ppi:left ?gene; ppi:right ?validatedProtein.
        } UNION {
                ?biogridInteraction ppi:right ?gene; ppi:left ?validatedProtein.
        }
?validatedProtein ppi:officialSymbol ?validatedProteinName .
OPTIONAL {?biogridInteraction ppi:published_in ?publication}.
} LIMIT 100
```

Listing 4.6: Query demonstrating use case 1.

4.3.3.2 Use Case 2: Domain-domain interactions candidates

Protein-protein interactions can be identified experimentally through various types of experiments (e.g: Yeast Two-Hybrid). However, it is not possible to identify the interacting domains between two proteins with the same experiments and it requires a set of different protocols and experiments. Often protein domains act as signature elements and repeatedly interact with each other within the same organism. Therefore these frequent observations assist in the identification of novel domain-domain interactions which helps the discovery of latent PPIs. Nevertheless, in this work such observations are inferred implicitly from the validated PPI dataset (BioGRID) and require further analysis to determine statistical significance. The SPARQL query example given in listing 4.7, retrieves the validated and the inferred scores for the domain interactions in which the HLH domain (*pfam:PF00010*) is involved.

4.3.3.3 Use Case 3: Selective interactions between segments of the Human genome.

Human chromosomes have a very compact structure in three-dimensional space and each chromosome appears to be folding into its own territory [151]. Even though the exact

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX pfam: <http://linkedlifedata.com/resource/pfam/>
 SELECT DISTINCT ?desiredDomainName ?interactingDomain ?interactingDomainName
   ?validatedScore ?inferredScore WHERE {
        pfam:PF00010 ppi:officialSymbol ?desiredDomainName .
        {
                ?interaction ppi:left pfam:PF00010; ppi:right ?interactingDomain
        } UNION {
                ?interaction ppi:right pfam:PF00010; ppi:left ?interactingDomain
        }
        ?interactingDomain ppi:officialSymbol ?interactingDomainName .
        OPTIONAL {?interaction ppi:score ?validatedScore} .
        OPTIONAL {?interaction ppi:inferredScore ?inferredScore}.
}
```

LISTING 4.7: Query demonstrating use case 2.

relation of spatial conformation of genes and their functionality is not fully understood yet, studies have shown that the structure of the human genome follows its functionality [158]. It is widely believed that chromosomal folding brings functional elements into close proximity regardless of their inter- or intra-chromosomal distance in base pair unit. In other words, the concept of close and far distances in relation to the spatial map of the genome is represented differently. Also, it has been shown that the likelihood of interactions between small and gene-rich chromosomes are more frequent [151]. All of these evidence suggest that the chromosomal conformation has implications for the gene activity and the functionality of its products, i.e. the proteins. Identifying the significance of the correlation between the genomic location of genes on one side and the selection of interaction partners for the genes products (the proteins) on the other side, will help to better understand the proximity patterns and the consequences from this arrangement. The prospective patterns can then be embedded into the prediction models in order to infer potential protein interactions.

As it has been described in the previous sections, the boundaries of ideogram bands on each chromosome have been selected as a metric for the genomic location for each gene. A number of genes reside on one ideogram as well as there are genes that reside between two consecutive ideograms. The reason for this selection decision is to facilitate the identification of neighbouring genes, their co-expression and potentially shared biomolecular mechanisms. The protein pairs from the PPI dataset have been mapped to their genomic location and their loci-loci association scores have been calculated based on the scoring function which has been described in section 3.2.3.4. Based on these findings, LinkedPPI provides the genomic location pattern in which proteins preferentially selected their interacting partners. These regions contain genes involved in the same pathways or share the same functionality which are yet to be identified by further gene enrichment analysis. The example SPARQL query provided in the listing 4.8 retrieves the pre-calculated inferred score between two ideograms, e.g. *Chromosome 3-q29* and

```
PREFIX ppi: <http://data.bioinfo.deri.ie/>
PREFIX entrezgene: <http://linkedlifedata.com/resource/entrez-gene/>
SELECT   DISTINCT ?proteinNameA  ?ideogramA ?proteinNameB ?ideogramB
?inferredScore ?interaction WHERE {
entrezgene:3280 ppi:officialSymbol ?proteinNameA .
?uniprotA ppi:entrezGene entrezgene:3280.
?uniprotA ppi:Ensembl ?ensA.
?ideogramA ppi:partOfIdeogram ?ensA.

entrezgene:23411 ppi:officialSymbol ?proteinNameB .
?uniprotB ppi:entrezGene entrezgene:23411.
?uniprotB ppi:Ensembl ?ensB.
?ideogramB ppi:partOfIdeogram ?ensB.
{
        ?interaction ppi:left ?ideogramA ; ppi:right ?ideogramB
} UNION {
        ?interaction ppi:left ?ideogramB ; ppi:right ?ideogramA
        }
?interaction ppi:inferredScore ?inferredScore.

}
```

LISTING 4.8: Query demonstrating use case 3.

Chromosome 10-q24.32 containing the protein-coding genes for the proteins HES1 and SIRT1 (*entrezgene:23411*), respectively.

4.3.4 Domain-domain interaction dataset

The example given in section 4.3.3.2 presented the retrieval of the domain partners for a known domain. However, by mapping the protein interaction partners to their domain composition a new set of domain-domain interactions has been extracted from the LinkedPPI consolidated dataset. From the comparison between the list of domain pairs from the 3did database with the selection of domain-domain interactions inferred from the PPIs in BioGRID through the use of LinkedPPI, a total number of 4'913 binary pairs of domain-domain interactions have been identified as potential novel interacting domain pairs. This dataset of domain-domain interactions is accessible from the link below [7].

Domain pairs which are reported in the new dataset have been extracted from pairs of interacting proteins with one domain. The inference hase been done based on at least a single occurrence in LinkedPPI, i.e. regardless of their frequency or likelihood of co-occurrence in the data repository. As a result, domain interactions at low frequency are not over-shadowed by frequently observed domain interactions which have been derived from well-studied proteins.

[7] goo.gl/eWGvXi

4.4 Discussion

The incorporation of complementary datasets for the expansion of PPI networks is a useful approach to gain insight into biological processes and to discover novel PPIs which have not been documented in the current PPI databases. However, there is an inherent high level of heterogeneity at the schema and instance level of these data sources, due to the lack of a shared schema for the representation of the data. Hence, Linked Data principles were applied for the data integration, retrieval and visualisation of the data to retrieve concealed knowledge.

The employment of the Linked Data principles is linked to the application of domain-specific models which can accommodate the needs in the field of PPI modelling and avoid the complexity of comprehensive ontologies. The use of a domain-specific model and an interactive graph-based exploration platform for search and aggregated visualisation makes this integration approach more intuitive for the domain experts. This chapter demonstrated a set of three use case scenarios depicting how the LinkedPPI framework could be used for the extraction of interaction candidates between proteins, domains and genomic regions. Nevertheless, the use of the LinkedPPI SPARQL endpoint is not limited to these three use cases only.

The SPARQL endpoint of LinkedPPI offers full freedom in articulating queries that are solely based on the user's research questions. However, the user profits from a graph-based visualisation interface, and the graph-based visualisation can still benefit from further advancements and alternative use cases. Currently, the queries would only produce direct interactions between proteins, domains or genomic loci in the graph representation, however, it could be advantageous to expose further hops in the interaction network. This way users could gain a better insight on the downstream and upstream elements in a more network like representation of the chain of events. On the other hand, it leads to a more congested network representation.

In addition, the current graph-based representation focuses on the single entry according to the user's query. However, the interplay between multiple components of a network is more instructive. Thus, it would be more advantageous for the domain users if a multi-entity query could be executed and then rendered with LinkedPPI instead of using the SPARQL endpoint of LinkedPPI for such a query.

The approach presented in this work allows the extraction of valuable information with regard to the PPI network, domain-domain interactions and selective genomic interactions. However, the observations reported in the outcome of such data retrieval is raw and could be a valuable asset for simulations and prediction methods, if the results are

fed into a procedure where further statistical analysis can be carried out. Thus, the integration of statistical analysis methods for the identification of significance and relevance of observations has to be considered for future work.

The integration of co-expressed genes in LinkedPPI is solely based on RDFized data provided by COXPRESdb, however, mRNA expression data is an expedient source in such integrative approaches. Therefore the integration of mRNA expression data from different tissues and cell lines from the EBI [8] will have to be brought into any extension of LinkedPPI as future work as well. This would expand the application of LinkedPPI toward other aspects of tissue specific pathway analysis.

Even though the proteomics data contains a high rate of false positive data, the tissue specific pathway analysis can largely benefit from an additional dimension that proteomics data supplies. Hence LinkedPPI can largely benefit from the incorporation of the data provided by Human Protein Atlas [9] in the future.

Human Protein Atlas is an extensive and complete source for protein expression in various cell lines, tissues and cellular compartment. Human Protein Atlas reports both mRNA and protein expression level which creates yet another way to narrow down potential protein partners for a protein in question. Hence, future RDFization and integration of the Human Protein Atlas into the current LinkedPPI will add another valuable layer of information. To achieve this a data model which can capture the relationship of different data types in both the Human Protein Atlas and the LinkedPPI dataset is needed. Future work will focus on development of such a data model as well as incorporating the data into the current visualisation platform.

[8] http://www.ebi.ac.uk/gxa/home
[9] http://www.proteinatlas.org/

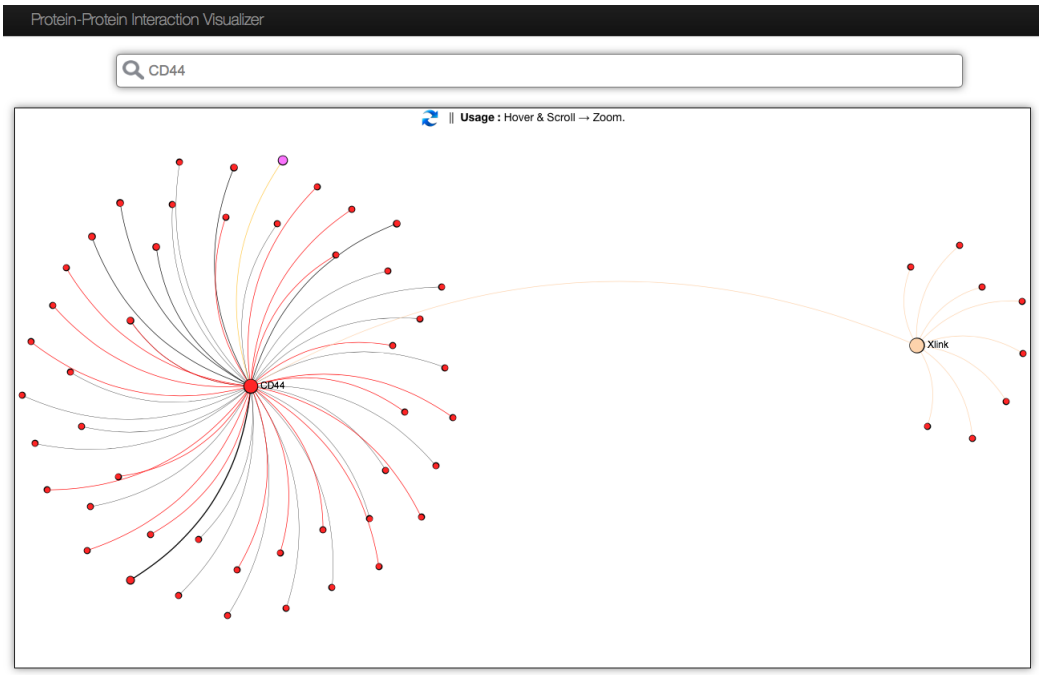
Chapter 5

Experimental Validation

This chapter reports on the experimental validation and biological findings. The hypotheses which have been under investigation in this chapter are driven from the computational models in the two preceding chapters. In this chapter the interaction between multiple proteins in mesenchymal stem cells (MSCs) isolated from bone marrow of mice and the role of these interplay in modulating the chondrogenic differentiation in MSCs is demonstrated.

In recent years, the application of MSCs in cellular therapy received very high attention with regards to the treatment of selected disorders. A considerable number of research experiments have been conducted in order to better understand the molecular pathology of joint related ailments such as osteoarthritis (OA). OA is characterised by cartilage damage, joint erosion and dysfunction. Thus it is vital to understand the network of protein interactions that cause the damage as well as the repair mechanisms. OA is tightly coupled with the degradation of the chondrocytes' extra cellular matrix (ECM); therefore this chapter is focused on components of ECM which are involved in the synthesis and stability of the cartilage.

As previously explained in the introduction to chapter three ECM proteins, CD44, TSG-6 and HAPLN1 are selected in order to validate the accuracy of the LinkedPPI platform. It is noteworthy that the choice of MSCs and selection of proteins were solely based on the expertise and interests of the laboratory. However, in theory the proposed model could be applied to any protein and its potential partners. Fig. 5.1 illustrates the network which is generated by LinkdPPI if the entry protein is CD44. The proteins and genes are shown in red circles. Experimentally validated interactions are shown by black links and the co-expression genes with the Pearson correlation coefficient are shown by red links. Protein names and gene symbols are not shown in this network for simplicity purpose. According to the network CD44's known protein domain is *Xlink*



Red nodes: proteins and genes. Light Brown: protein domain. Purple nodes: Cellular components. Black links: experimentally validated. Red links: co-expressed genes.

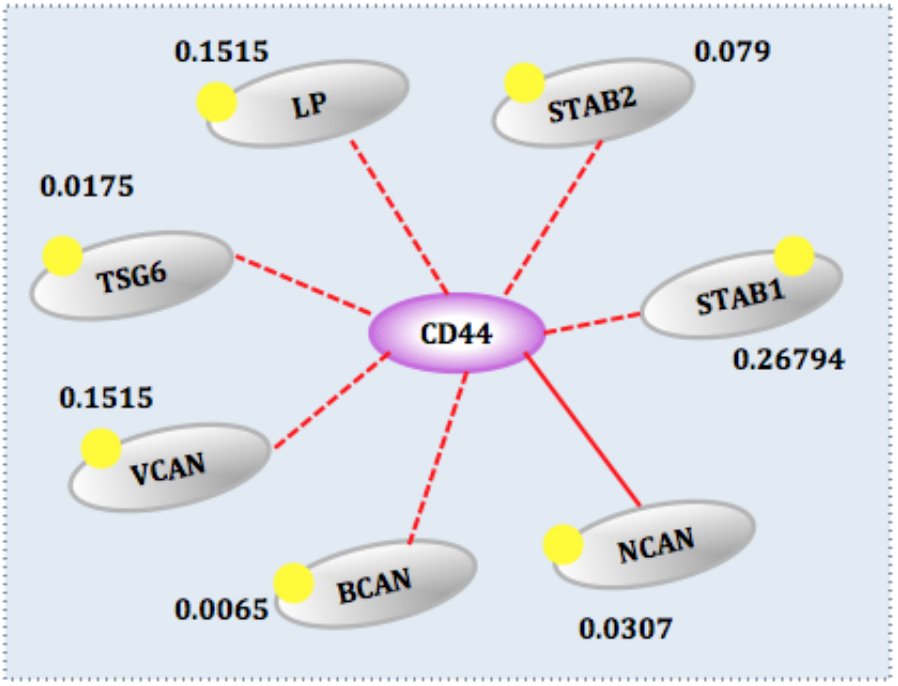
FIGURE 5.1: LinkedPPI output for CD44 search.

or *link* domain which is observed in 8 other proteins too. Since the objective of this chapter is to investigate the effect of protein's domain composition and their genomic location on deducing novel interacting partners, only the sub part of this network which contains the domain information is shown in Fig. 5.2.

Fig. 5.2 illustrates a network of potential protein partners for CD44. This network is expanded from Fig. 5.1 and the loci-loci interaction score which was extensively explained in Chapter 3 were retrieved for each protein from the LinkedPPI SPARQL endpoint. As it is shown in the figure each of these proteins contain a 'link' domain and is therefore a member of the HA binding protein family. However, the loci-loci score varies for each protein. The highest loci-loci score is shown for STAB1 and VCAN. These two proteins have a similar domain composition to TSG-6. However, STAB1 and VCAN have a molecular weight of 275 and 372 kDa respectively which make them large molecules for protein detection. Therefore, they are not the most suitable candidates for such exploration. Hence, TSG-6 and HAPLNA1 (LP) were the two proteins which have been selected based on domain expert knowledge and interests.

In summary, the relevant proteins for this study are as follow: (i) CD44 which is a cell surface protein, (ii) Tumor Necrosis Factor-Stimulated Gene-6 Protein (TNFAIP6) a secreted protein, and (iii) Cartilage-Linking protein 1 (Crtl1) also known as Hyaluronan

and Proteoglycan Link Protein 1 (HAPLN1), a member of the extracellular cartilage matrix.

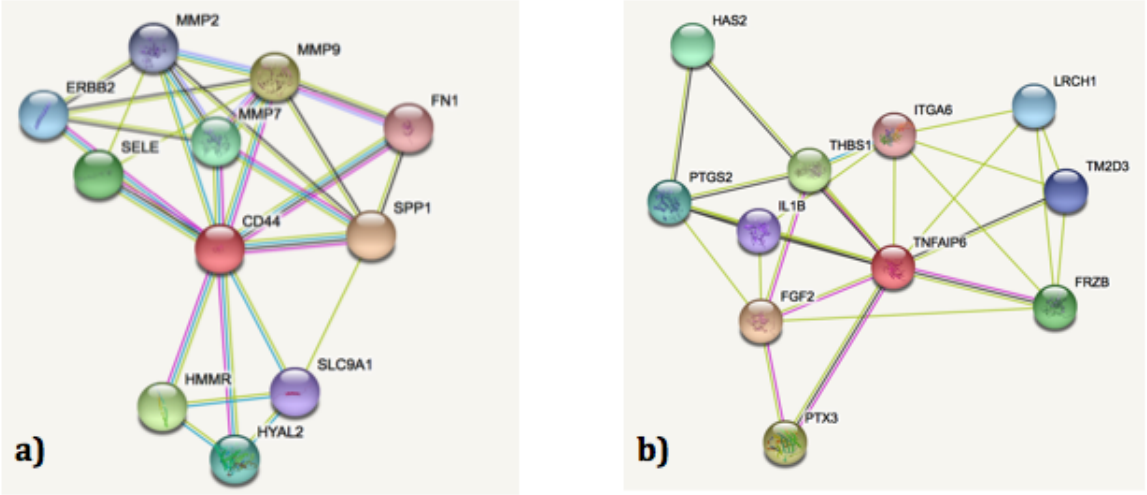


Dashed line: predicted interactions. Solid line: experimentally validted. Yellow circle: Link domain. Loci-loci score is shown for each partner.

FIGURE 5.2: CD44 partners retrvied form LinkedPPI.

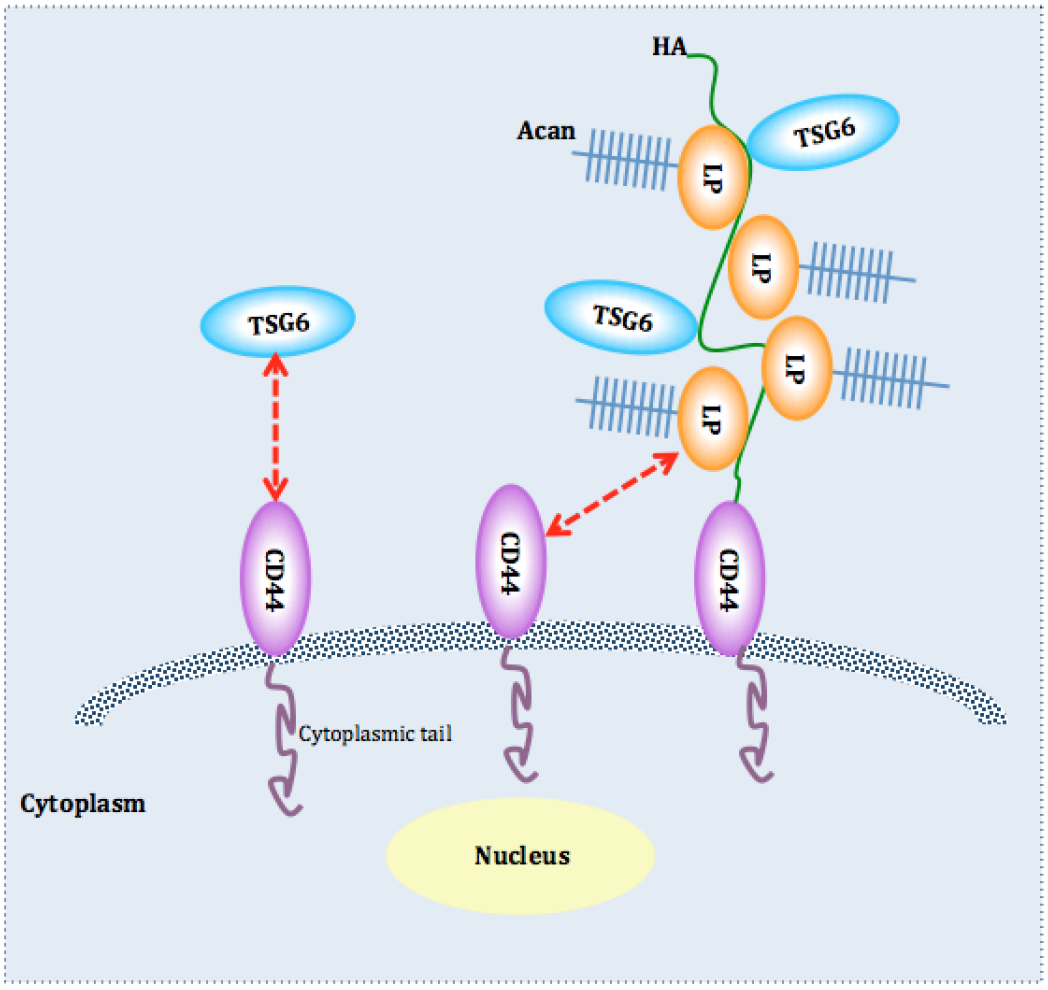
5.1 Overview of the interaction network

As explained in the introduction chapter, the CD44, TSG-6 and HAPLN1 proteins have shared commonalities including their involvement in cellular processes such as cell migration, inflammation response, HA-binding affinity, maintaining cartilage ECM stability and development of OA. Thus, this chapter investigates on the relation between these proteins. Figure. 5.2 depicts the network retrieved based on LinkedPPI visualisation while figure. 5.3 illustrates the predicted protein partners for both CD44, (a), and TSG-6, (b), based on String database [139]. As can be seen neither of the proteins were identified as each others partner in the String prediction. Therefore, make a good example which demonstrates the potential of LinkedPPI in identification of novel protein partners. Figure. 5.4 illustrates the hypothetical network of interactions between CD44, TSG-6 and HAPLN1.



a) possible protein partners of CD44. **b)** possible protein partners of TSG-6.

FIGURE 5.3: Network of interactions for CD44 and TSG-6 predicted by String.



HA: Hayluronan. **LP:** HAPLN1. **Acan:** Aggrecan. Dashed arrow: predicted interactions.

FIGURE 5.4: Network of interactions between CD44, TSG-6, HAPLN1.

5.2 Methods and Materials

5.2.1 ATDC5 cell culture and expansion

ATDC5 cells, a mouse chondrogenic cellline, [159] were retrieved from the European Collection of Cell Cultures (ECACC). ATDC5 cells have been expanded in Dulbecco's Modified Eagle's Medium and Ham's F-12 Nutrient Mixture (DMEM/Hams F12) + Glutamine (Sigma), 5% Fetal Bovine Serum (FBS) (Thermo Fischer) with 1% penicillin/streptomycin (Gibco). Upon reaching 70-80% confluence, ATDC5 were subcultured by washing them with 5ml of D-PBS which has then been discarded. Subsequently, 5ml of 0.25% trypsin/ethylenediamine tetraacetic acid (EDTA) (Gibco) has been added to each T-175 flask and left to incubate at for 5 minutes. To neutralise the 0.25% trypsin/EDTA, 10ml of ATDC5 expansion medium was added to each flask. The ATDC5 suspension was then transferred to a 50ml sterile centrifuge tube. The ATDC5 cell suspension was then centrifuged at 400G for 5 minutes. The cell supernatant was then removed and discarded, and the cell pellet resuspended in 10ml of ATDC5 expansion medium and a cell count has been performed using a haemocytometer. ATDC5 were plated into T-175 flasks at a density of $1x10^6$ cells per unit of cm^2.

5.2.1.1 Chondrogenic differentiation

ATDC5 cells were seeded at 2.2 million cells per T-175 flask at a density of Ĩ2,500 cells/cm^2, and the cells remained in expansion medium for 24 hours prior to differentiation. 24 hours after initial seeding, the medium was changed from expansion to differentiation medium. Differentiation medium consisted of DMEM high glucose (Sigma), 100nM dexamethasone, 50 μg/ml ascorbic acid 2-phosphate, 40 μg/ml L-proline, 1% Insulin-Transferrin-Selenium-Ethanolamine (ITS) supplement, 1mM sodium pyruvate, 1% penicillin/streptomycin with 2% FBS. ATDC5 cells were differentiated in normoxia for up to 21 days.

5.2.1.2 Alcian Blue staining of ATDC5

Following differentiation, medium was removed and ATDC5 washed twice with D-PBS. Cells were fixed with 95% methanol for 30 minutes. and stained with 1% Alcian Blue over night at room temperature (RT). The Alcian Blue was removed and the flasks were washed several times with dH2O to remove excess stain. Positive blue staining indicating matrix proteoglycan production was then imaged using an Olympus BX51 microscope.

5.2.2 Detection of Proteins Expression

5.2.2.1 Cell Harvest

ATDC5 cells were harvested by removing medium and washing twice with ice-cold D-PBS. Cells were detached using trypsin and placed back into appropriate cell culture incubators (37℃) for 5 minutes. Trypsin neutralise using ATDC5 expansion medium and the cell suspension was transferred to 15ml tubes. Cells were centrifuged at 400G for 5 minutes at 4℃. Medium was removed and discarded and the cell pellet resuspended in 1ml of ice-cold D-PBS. Cells were then centrifuged again at 400G for 5 minutes at 4℃ and D-PBS was removed. The cell lysis buffer was prepared by adding phenylmethanesulfonylfluoride (PMSF) (0.3M) to a final concentration of 1mM, and protease inhibitor cocktail (10X) (Life Technologies) to a final concentration of 1X. The cell pellet for each time-point was lysed in 50 μl cell lysis buffer for 30 min and vortexed during this time at 3 minute intervals. This extract was spun at $\tilde{1}$5,100g (13,000 RPM) for 10 min at 4℃. Then the clear lysate was transferred to a clean tube and stored at -20℃.

5.2.2.2 Quantification of Proteins Using BCA Assay

Protein quantification was determined by bicinchoninic acid (BCA) assay (Thermo Scientific) according to the manufacturers instructions. A standard curve was created using the supplied albumin standard. Unknown samples and standards were added to the 96 well plate. A working reagent stock was created by combining 50 parts BCA reagent A and 1 part BCA reagent B. 200 μl of the working reagent was then added to each well and the plate incubated at 37℃ for 30 minutes. The absorbance of the samples and standards was read at 562nm on a Wallac Victor 3 plate reader.

5.2.2.3 List of Primary and Secondary Antibodies

- Rat monoclonal anti-CD44 raised against spleen cells from LOU/MN rats with a bone marrow-derived stromal clone, (Santa Cruze:sc-52536, 1:500).
- Mouse anti-TSG-6 (D-4) is a mouse monoclonal antibody specific for an epitope mapping between amino acids 22-51 at the N-terminus of TSG-6 of human origin, (Santa Cruze: sc-398307, 1:500).
- Goat anti-mouse HAPLN1 is an affinity purified goat polyclonal antibody raised against a peptide mapping within an internal region of HAPLN1 of human origin, (Santa Cruze: sc- 46826, 1:500).

- Goat anti-rat, HRP, (Santa Cruze: sc- 2006, 1:1000).
- Goat anti-mouse IgG, (CalBioCHEM, # 401215, 1:1000).
- Rabbit anti-goat IgG, (CalBioCHEM, # 401504, 1:1000).
- Donkey Anti-Goat IgG HL (TRITC),(abcam, #ab6882,1:800).
- Goat Anti-Rat IgG HL (Alexa Fluor 488),(abcam, #ab150157, 1:800).
- Anti-Mouse IgG (H+L), F(ab)2 Fragment (Alexa Fluor 555 Conjugate), (Cell Signalling Technology, #4409, 1:800).
- Mouse IgG isotope control, (BD Pharmingen, cat : # 556648, 1:1000).
- Rat IgG isotope control, (BD Pharmingen, cat : # 559072, 1:1000).

5.2.2.4 Western blot (WB)

Western blotting was performed as follow: Samples were prepared by adding 20 μg protein/well, loading buffer (4X) (Thermo Scientific - Composition: 40% glycerol, 4% lithium dodecyl sulfate (LDS), 4% FicollTM 400, 0.8 M triethanolamine-Cl pH 7.6, 0.025% phenol red, 0.025% coomassie G250, 2mM EDTA disodium) with β-mercaptoethanol (16%) and dH2O to a final volume of 20μl. The samples have been heated at 95℃ for five minutes and given a short spin in the centrifuge. A 10% acrylamide pre-cast gel (Bio-Rad) was run at 110V for 1hr 45min in running buffer (1x). The polyvinylidene fluoride (PVDF) membrane (GE Healthcare) was immersed in 100% methanol for 10 min, deionized water (dH2O) for 3 min and transfer buffer (1x) for 3 min to activate it. The transfer was set up and run at 100V for 1hr 30min in transfer buffer (1x) at 4℃. The membrane was rinsed with TBS-Tween (TBST) (0.05%). A milk blocking solution (1%) was prepared and the membrane incubated in this solution for 1hr at RT. The membrane was incubated with primary antibodies overnight at 4℃ on a roller. The membrane was then washed 3 times for 5 min in TBST. The membrane was incubated with secondary antibody for 1 hr at RT on a rocker. The membrane was finally washed 3 times for 5 min in TBST.

5.2.3 Co-Immunoprecipitation (Co-IP)

Pre-clearing phase

A minimum of 500μg of lysate was diluted to 500μl with normal lysate buffer which has been described in section 5.2.2.1. Next 1μl of an irrelevant antibody of the same

TABLE 5.1: Western blot buffers.

Buffers	Reagents
Running Buffer 10X	Glycine 144g
	TrisBase 30.35g
	SDS 10g
	Up to 1L dH2O
Transfer Buffer 10X	Glycine 144g
	TrisBase 30.35g
	Up to 1L dH2O
Wash Buffer 10X (TBS)	TrisBase 121g
	NaCl 40g
	Adjust to pH 7.6
	Up to 1L dH2O

host species was added and incubated at 4℃ with rotation for 1 hour. After 1 hour 50μl of protein A Agarose was added and incubated at 4℃ with rotation for 30 minutes. Then the mixture was centrifuged at 3000g for 3 minutes. Next, the supernatant was transferred to a fresh tube for the immunoprecipitation steps. The agarose pellet has been added to 30μl of western blot loading buffer to run on the gel as the pre-clearing sample.

Immunoprecipitation phase

2μl of the primary antibody of the tagging protein was added to the pre-cleared lysates and incubated at 4℃ with rotation overnight. The next day 50μl of protein A agarose was added to the pre-cleared lysates and incubated at 4℃ with rotation for 1 hour. Then lysates was centrifuged at 3000g for 3 minutes at 4℃. The supernatant was discarded and the agarose pellet has been resuspended in 50μl of wetsren blot loading buffer. Western blot was performed according to section 5.2.2.4.

5.2.3.1 Protein visualisation by Enhanced Chemiluminescence

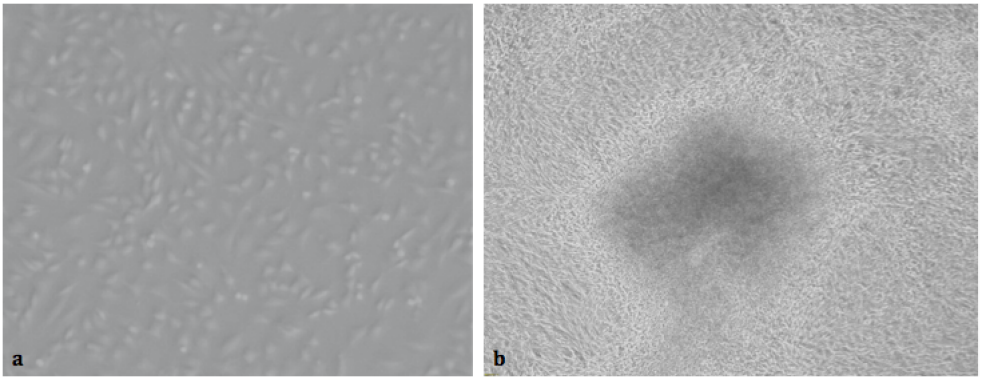
Visualisation of the proteins of interest was performed using electrochemiluminescence (ECL) detection reagents (GE Healthcare). The membrane was incubated in the ECL at a ratio of 1:1, reagent A: reagent B for 2-3 min. The membrane was been covered in saran wrap and placed in a dark box before imaging using an Alpha Innotech Fluorchem FC2 Multi-Image II.

5.2.3.2 Immunostaining of mono layer ATDC5

Slides were rinsed with PBS and fixed with 95% methanol for 30 minutes and was rinsed with PBS 3 times for 5 minutes at RT. Slideswere incubated for 60 minutes at 37 ℃ in 1%BSA and Tris Acetate buffer with Chondroitinase ABC (1:125) in order to break down the matrix structure. After one hour slides were rinsed with PBS for 3X5(min). Then slides have been blocked with 1%BSA and PBS buffer for 60 minutes at 37℃. After that slides were incubated with primary antibodies (1:1000) with 0.5% BSA and PBS buffer at 4℃ overnight in a humidity box. Next day slides were washed with PBS for 3 time for 5 minutes at RT. Then slides were incubated with the secondary antibody (1:800) and DAPI for staining the nucleus (1:2000) for one hour at RT in a dark box. Slides were rinsed with PBS for 3 times for 5 minutes and mounted with a coverslip in the dark box. Images have been captured using the Andor Revolution confocal microscope.

5.2.3.3 Immunostaining of knee sections

Slides were rehydrated with Xylene 2 times for 5 minutes, 100% ethanol 2 times for 3 minutes, 95% ethanol for 3 minutes and 70% ethanol 3 minutes. Slides were washed briefly in water and TBS. Slides were air dried and sections have been marked with the hydrophobic pen. Slides were incubated for antigen retrieval with 40μU Chondroitinase ABC and 0.1M Tris Acetate, 1% BSA for 30 minutes at RT. Then were been washed in TBST 3 times for 5 minutes. Then slides were incubated with blocking solution (TBS/Tween (0.05%) and 5% BSA) for 1.5 Hours at RT and rinsed in TBST 3 times for 3 minutes. Slides were incubated with primary antibodies solution: PBS/Tween (0.05%), 5% FBS and antibodies overnight at 4℃. Next day the primary antibody was washed by TBST 3 times for 5 minutes and incubated with fluorescent labelling solution (TBS/Tween (0.05%) and secondary antibodies) in the dark at RT for 1 hour. The labelling solution was rinsed in TBST for 3 times for 5 minutes. Finally, slides were incubated for dehydration with 70% ethanol-3 mins, 95% ethanol-3 minutes, 100% ethanol-2 times for 3 minutes, Xylene- 2 times for 5 minutes. Slides were mounted by adding 3 drops of mounting medium to the surface and placing the cover slip on the top of the section.



a) Expansion on day 0. **b)** Chondrogenic differentiation on day 21.

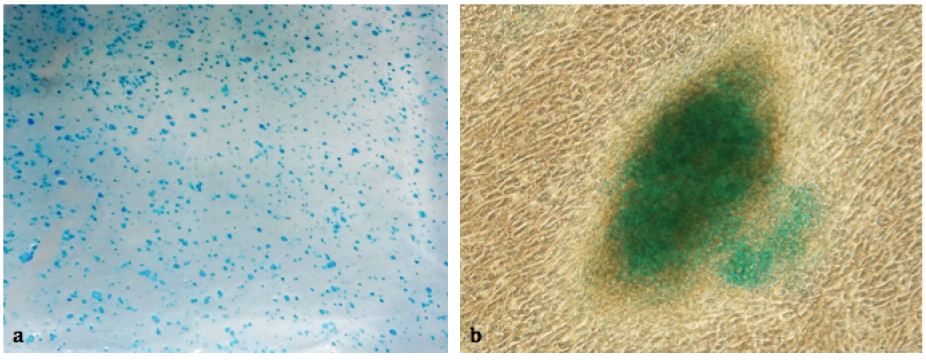
FIGURE 5.5: Monolayer cultured ATDC5 cells, 4x magnification.

5.3 Results

5.3.1 Morphology of ATDC5 cells in culture and chondrogenic differentiation

ATDC5 cells were cultured and passaged in order to obtain the required number of cells for protein quantification and qualification. ATDC5 cells were received at passage 3 and expanded till passage 9 whereupon chondrogenesis was inducted. Cells were cultured under two medium conditions: expansion and chondrogenic as described in the methods section. Pellet were harvested at the following time points: days 0, 3, 7, 14, 21 for protein quantification. Figure 5.5 displays the morphology of the ATDC5 cells on day 0 and day 21 during expansion and chondrogenesis.

Figure 5.4 shows the formation of chondrogenic nodules on the surface of flask and Alcian Blue staining for GAG in these nodules.



a) Chondrogenic nodules on the surface of the flask. **b)** GAG staining of a single nodule. 10x magnification.

FIGURE 5.6: Confirmation of chondrogenesis by Alcien blue staining for GAG on day 21.

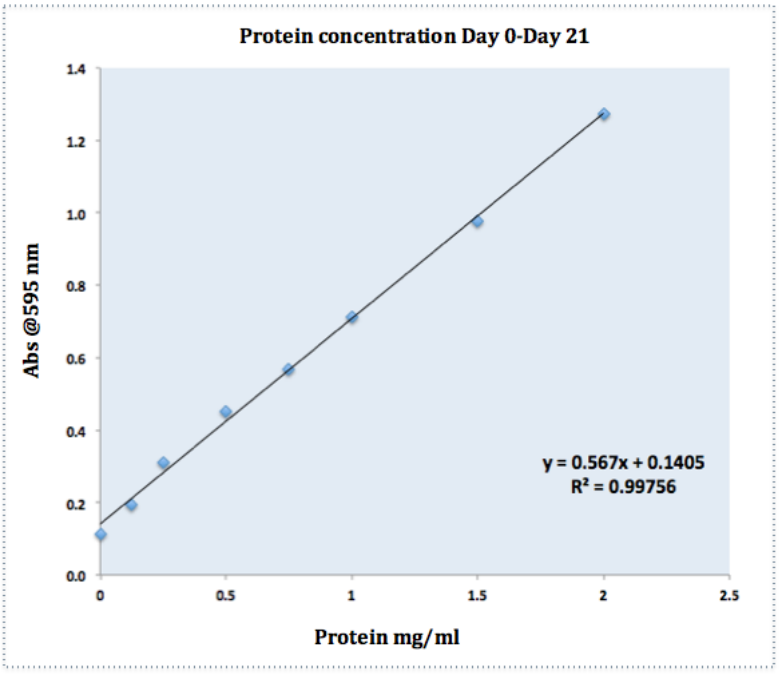


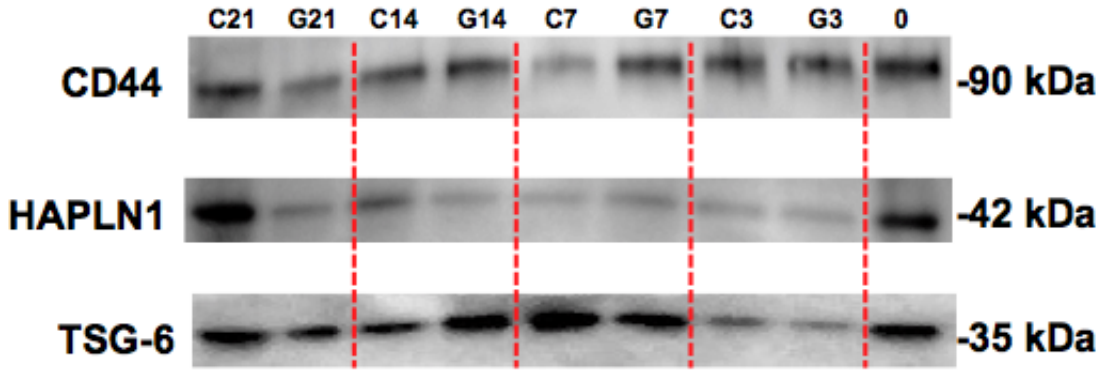
FIGURE 5.7: Standard curve generated to determine protein concentrations of ATDC5 cells at Day 0 to Day 21 in both expansion and chondrogenic media.

5.3.2 Quantification of protein concentration in ATDC5

Proteins from previously indicated time points for both expansion and chondrogenesis were isolated from cell pellet based on the lysis protocols which have been described in the methods section. Subsequently, the BCA assay was performed and the protein concentration (mg/ml) of the cell lysates from day 0 to day 21 time points measured using the standard curve generated in Fig. 5.7. The final protein concentration which were used for each sample in western blot are given in table 5.2.

TABLE 5.2: Protein concentration.

Sample	Concentration	Final Concentration (mg/ml)
Day 0	0.264	4.595486247
Day 3G	0.431	1.950018276
Day 3C	0.516	1.509635181
Day 7G	0.484	1.651107726
Day 7C	0.460	1.773909578
Day14G	0.635	1.146729705
Day 14C	0.685	1.040859478
Day 21G	0.614	1.19627077
Day 21C	0.779	0.887399678



Selcted time points are day 0, 3, 7, 14, 21. C: chondorgeneic medium. G: growth/expansion medium.

FIGURE 5.8: Expression of CD44, TSG-6, HAPLN1 proteins during expansion and chondogenesis.

5.3.3 Detection of protein expression in chondrogenesis

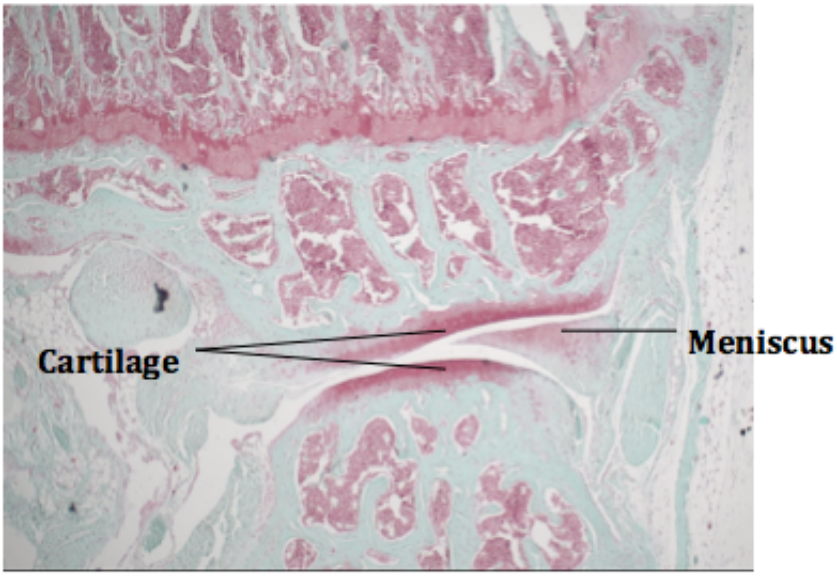
In order to identify whether CD44, TSG-6 and HAPLN1 proteins are expressed in ATDC5 cells, western blot was carried out. Fig. 5.8 shows the western blot images for the protein expression levels for selected time points in expansion and chondrogenic conditions. Using a 1:500 dilution of primary antibodies, bands were detected at 80kDa, 42kDa and 35kDa which are the expected molecular weight for CD44, HAPLN1 and TSG-6, respectively.

5.3.4 Detection of proteins co-localisation by immunofluorescence (IF)

In order to determine the co-localisation of pairs of proteins in the same cellular compartment, IF experiments were carried out on monolayer cultured ATDC5 cells as well as mouse knee sections.

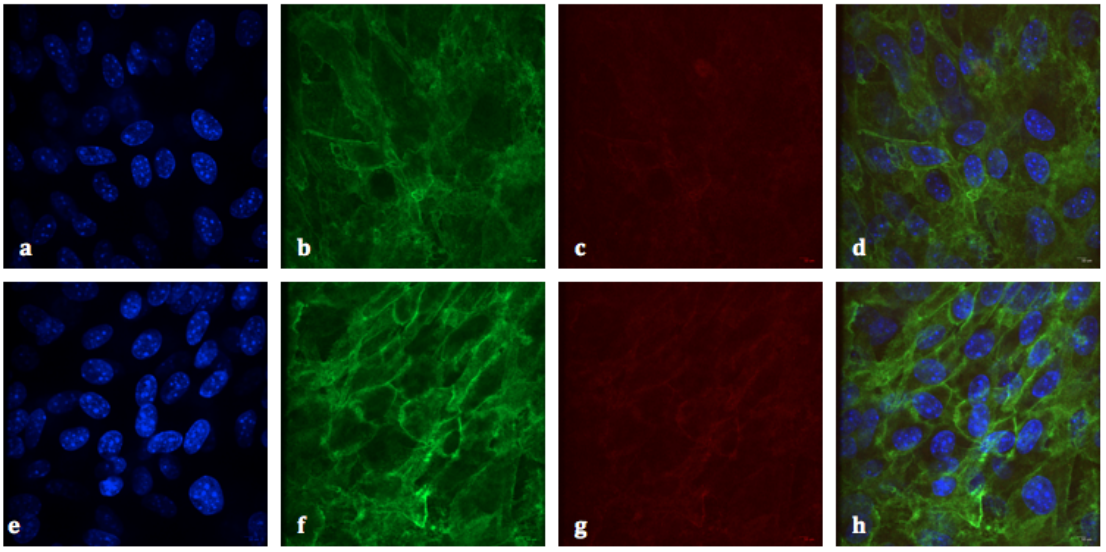
These sections were obtained from a mouse knee previously treated with collagenase to induce osteoarthritis change on control knee. Figure. 5.9 shows the Safranin O staining of the cartilage in the control section. Fig. 5.10 depicts the expression of CD44 (green), TSG-6 and HAPLN1 (red) with regards to the location of the nucleus (blue) in a monolayer on day 21. As can be seen, CD44 is expressed on the cell surface and in the ECM as expected. However, both TSG-6 and HAPLN1 do not show positive staining. The images have been taken with a confocal microscope at 60x magnification.

Fig. 5.11 displays the IF on mouse knee sections which have been treated by collagenase to induce OA. The control is shown in upper panel and OA affected is in lower panel. IF



Staining was performed by Dr. Niamh Fahy

FIGURE 5.9: Safranin O stained control knee section.

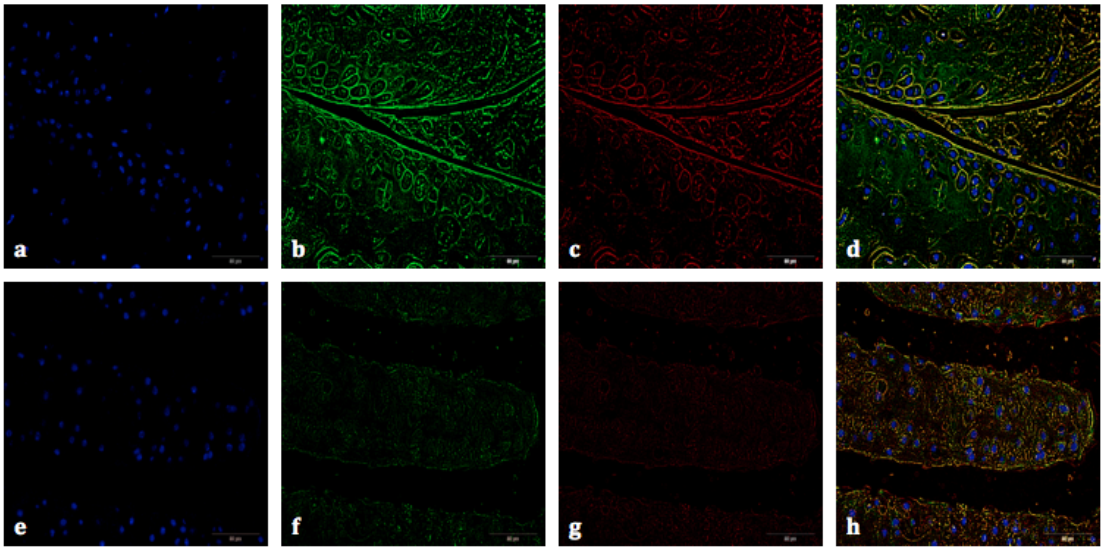


Upper panel: CD44-TSG-6, lower panel: CD44-HAPLN1.**a,e)** Nuclei blue. **b,f)** CD44 green. **c)** TSG-6 red. **g)** HAPLN1 red. **d,h)** merged.

FIGURE 5.10: IF of ATDC5 culture monolayer on day 21, 60x magnification.

on knee sections has been carried out only for the co-localisation of CD44 and TSG-6. These sections illustrate the femur bone on the top, meniscus between the two bones and the tibia bone at the bottom. Staining of CD44 and TSG-6 are both highly visible. Moreover the image demonstrates apparent co-localisation of the pairs.

With respect to the IF image the OA affected knee section displays less obvious staining compare to the control section in Fig. 5.11. Since these sections are derived from OA



Upper panel: CD44-TSG-6 in control animal, lower panel: CD44-TSG-6 in OA affected animal. **a,e)** Nuclei blue. **b,f)** CD44 green. **c,g)** TSG-6 red. **d,h)** merged.

FIGURE 5.11: IF of collogenase treated mice knee section, 40x magnification, scale bar 50mM.

knees it is expected that the structure of the cartilage is disturbed thus co-localisation of the pair might be compromised in diseased condition.

5.3.5 Detection of protein binding

Subsequent to the identification of protein co-localisation with IF, co-immunoprecipitation (Co-IP) was performed in order to definitively determine whether these pairs bind to each other and appear in the same complex on day 21 of chondrogenesis. For this purpose Co-IP was carried out on the same protein lysis which has been used for the western blot previously. The Co-IP protocol was described in detail in the methods section. Fig. 5.12 displays blots for the Co-IP of CD44 and anti TSG-6 (a), CD44 and anti HAPLN1 (b) on day 0. The faded line of lane IP in blot (a) indicates the bindings between CD44 and TSG-6 on day 0. However, in the blot shown in (b) there is no visible line in the expected range for HAPLN1 which suggests there is no binding between CD44 and HAPLN1 on day 0.

Fig. 5.13 shows the Co-IP blots for CD44 anti TSG-6 on day 21 for both chondrogenisis and expansion conditions. As can be seen in the both images, clear bands appear around 35 kDa which is the molecular weight of TSG-6. The bands are more intense during chondrogenesis in contrast to the control conditions.

Fig. 5.14 shows the Co-IP blots for CD44 anti HAPLN1 on day 21 in both chondrogenisis and expansion conditions. In both blots appearances of the band in the IP lanes indicates

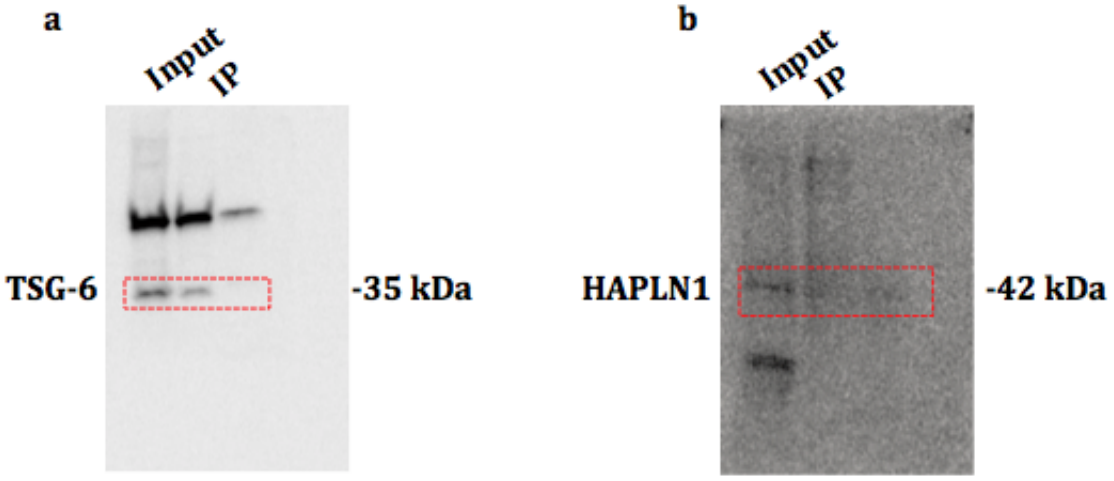
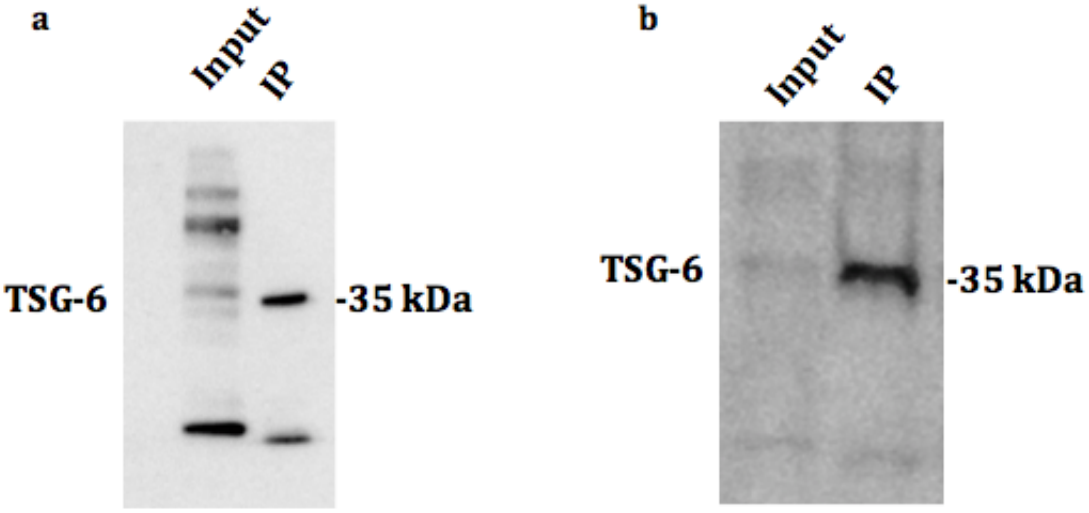


FIGURE 5.12: Co-IP of CD44 anti TSG-6 and anti HAPLN1 on day 0.

the interaction between CD44 and HAPLN1 protein. However, in order to determine the intensity of the band further densitometry analysis is required.

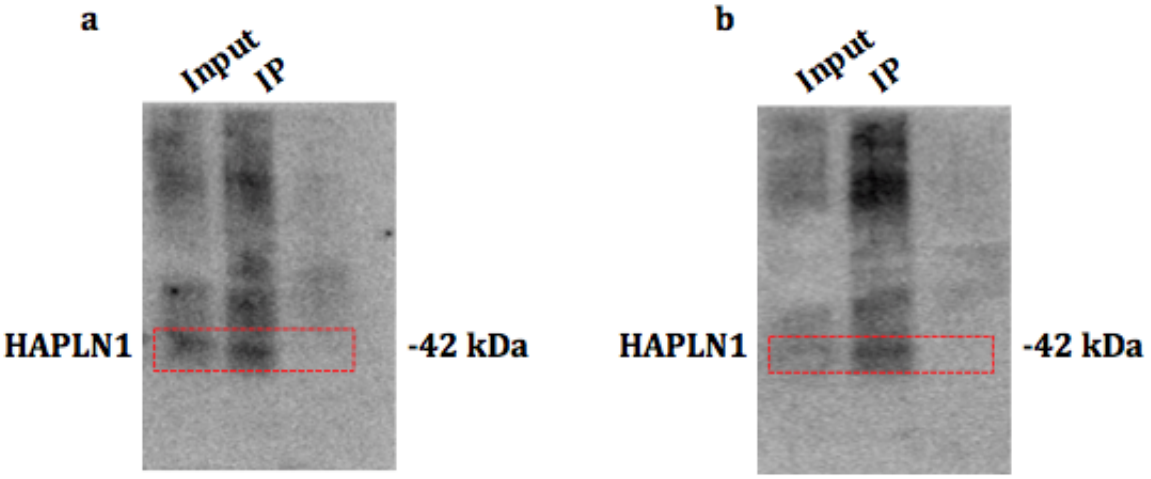


a) Chondrogenisis. **b)** Expanssion.

FIGURE 5.13: Co-IP of CD44 anti TSG-6 on day 21.

5.4 Discussion

The results which have been reported in this chapter demonstrate the physical association of CD44 and TSG-6 proteins in ATDCS5 cells. Considering the high importance of



a) Chondrogenisis. **b)** Expanssion.

FIGURE 5.14: Co-IP of CD44 anti HAPLN1 on day 21.

MSCs in cell therapy and in particular in finding treatments for diseases such as OA, it is essential to fully understand the underlying molecular mechanism of action in MSCs. Thus it is important to identify the crucial players their roles and their interplay.

Previously it has been reported that both CD44 and TSG-6 molecules have the ability to bind to HA molecules [160]. Additionally, they both are identified as interactors of aggrecan which is the most predominant glycan in the ECM cartilage. Thus it is interesting to investigate the strength of co-expression of CD44 and TSG-6. The regulation of CD44, TSG-6, HAPLN1 under chondrogenic conditions has been further confirmed by examining their proteins expression in ATDC5 cells differentiated for 21 days.

According to the results presented in Fig. 5.8 all these three proteins are highly expressed on day 0. However on Day 3, the expression level of CD44 and HAPLN1 distinctly dropped even though TSG-6 expression was still prominent across all time points to day 21. It is noteworthy that bands related to the growth condition appeared strongly in comparison to reference for chondrogenic conditions on day 3, 7 and 14. However, all these three proteins are boldly expressed on day 21 and in all the cases the chondrogenic bands are stronger than growth conditions. This fluctuation in the expression of CD44 and HAPLN1 proteins is an indication of a switch-like mechanism which suppresses their expression at the beginning of chondrogenic induction. The increase in expression levels of both CD44 and HAPLN1 on day 21 suggests this suppression mechanism has been overcome. The suppressive mechanism hypothesis also brings to mind the activation of a critical condition by either CD44 or HAPLN1 which requires the suppressive mechanism. Further, experiments are needed in order to verify this hypothesis.

As a complementary method to the western blot, the IF experiment has been performed on the ATDC5 cells for the identification of localisation of CD44, TSG-6 and HAPLN1, where the ATDC5 cells have been cultured in monolayer on chamber slides for 21 days. As it has been shown in the resulting images, the CD44 staining is visible however the stain for TSG-6 and HAPLN1 are not strong and clear. Therefore the co-localisation of CD44-TSG-6 in the cartilage of both healthy and OA knee joint has been investigated. Based on the images presented in Fig. 5.11 these two proteins appear to be highly co-localised.

Eventually, the western blot showed strong expression of CD44, TSG-6 and HAPLN1, but pairs of the proteins have not shown signs of co-localisation in monolayer cultured ATDC5 cells. Due to their strong association in the cartilage tissue, it can be concluded that their localisation and further interaction may accrue at a later stage in the chondrogenic process. The author, the direct physical interaction between CD44 and TSG-6 has not been previously reported in the literature.

In this chapter, the direct binding of CD44 and TSG-6 proteins was reported. However, despite the unchanging expression of TSG-6 over the course of chondrogenesis expression the fluctuation in the expression of CD44 before Day 21 points to an underlying unknown mechanism or element(s). A hypothetical mechanism or element(s) which might be the cause for this fluctuation by suppressing either CD44 or one of its regulatory elements. These possible mechanisms could be further investigated in future work.

In this research work, the interaction between CD44-TSG-6 has been confirmed only in mouse ATDC cells however it is necessary to confirm this binding in human MSCs as well. Thus further steps will include performing similar experiments on human MSCs.

Chapter 6

Conclusion

6.1 Thesis summary

In the advent of *'-omics'* technology the rate of data generation has been amplified to an unprecedented degree and is not anywhere near an upper threshold. Thus, procuring more advanced, accelerated and meticulous data analysis methods is inevitable and necessary. However, algorithms and methods are only part of data analysis and to a very high extent the data representation significantly influences the analysis. Hence, it is essential to achieve the highest veracity of what the deluge of biomedical data offers.

The *'-omics'* data is only a general term applied to the data which are produced by using the *'-omics'* technology, however, the differences in these data are immense as opposed to their scarce similarities. Nonetheless, these sparse similarities are adequate for linking the disparate biomedical data. Therefore, highlighting the analogies in two or multiple data sources and coupling them is the key to extend our accessible knowledge in particular biomedical research question.

In order to underline the sameness in multiple data sources, the data has to be represented and modelled in a standardised manner. However, despite the number of public efforts, data standardisation in the biomedical domain has not been fully accomplished. Nevertheless, the technologies based on Semantic Web principles offer data standardisation through the application of Linked Data principles. This thesis explored the relevance of a Linked Data approach applied to semantically modelling biomedical data for the prediction of PPIs.

The first objective of this thesis was to identify characteristic protein parameters that have the potential of being used as predictive features in a PPI network analysis. Similar to the choice of the right predictive approach which can influence the performance of

the predictive model, the set of features which have been used in the prediction alter the prediction outcome. Thus, it is essential to choose protein features which contribute towards more accurate predictions. The aim was to test innovative features by measuring their performance in models against the currently available PPI datasets. The prediction scores have contributed to the selection of protein interaction candidates that have to be validated in laboratory experiments. The detailed analysis that has been carried out has been presented in Chapter 3.

The second objective of this thesis was to demonstrate the efficiency of the Linked Data principles in the integration of multiple biomedical data sources with the aim to create a unified set of data from heterogeneous biomedical data sources. This thesis showed the advantages of the Linked Data principles and addressed its limitations in Chapter 4. The prediction scores from Chapter 3 have been provided in combination with the reference protein interaction data to guide users in the selection of protein interaction candidates.

The third objective of this thesis was to verify the findings from the two previous objectives and substantiate them with experimental methods in the laboratory. A variety of computational models was proposed which demonstrated the ability of the diverse range of data sources and algorithms in the prediction of PPIs however very few predictive models or features were experimentally validated. This thesis provided evidence on the potential of the newly identified protein features and their semantic integration in the prediction of PPIs by validating novel predictions with experimental methods in the laboratory. A thorough description of the methods and outcomes have been presented in Chapter 5.

6.2 Contributions

The core contributions of this thesis are summarised as follow:

Integration of features from the protein's 3D genomic location as a means to ground PPI predictions on genomic data. A prediction model was presented which has demonstrated the relevance of the genomic location for the prediction of the proteins' functionality and preferential bindings. The basic features for each pair of genomics loci was derived from the existing reference dataset for PPIs. Using these features in combination with other features (and with a Naïve Bayes approach) the relevance of loci-loci associations in contrast to the remaining features was demonstrated. The evaluation showed that this model outperformed the other state-of-the-art models.

Development of a domain specific schema for representing PPI network data (according to Linked Data principles). The domain specific schema serves as a basic ontology and enabled the transformation of the multi-dimensional data into RDF in order to maintain the interoperability and openness of all data sources. The unified data is publicly accessible through a SPARQL endpoint.

Graph-based visualisation platform for intuitive exploration of PPIs and related data. In this thesis, a graph-based visualisation platform for domain experts with no or less technical knowledge was proposed.

Domain-domain interaction dataset. A set of 4,913 potential domain-domain interactions has been extracted from the multidimensional data set which is presented in Chapter 3.

Validation of specific protein bindings in ECM of MSCs. The physical binding between two proteins residing on the cell surface and secreted from the cell to the ECM of MSCs in mice was investigated and their different level of expression under different conditions was shown.

6.3 Future research direction

In Chapter 3 the loci association has been analysed only on the datasets derived from human PPIs. The association pointed towards a shared underlying mechanism between proteins with similar or close functionality, therefore the existence of an operon like concept in eukaryotes is speculated. However, using merely human data is not sufficient to draw a certain conclusion. Hence, a similar study needs to be done on other model organism like mice, rat and yeast for which abundant volume of data is available in order to be able to establish the possibility of the presence of such a phenomena. Nonetheless studying other model organisms may not suggest similar observations hence investigation of the evolutionary diverged from this hypothesis will shed light on the differences and causation of this deviation.

Looking into the preferential interactions between genomic loci in other model organisms follows the same concept as human, therefore a similar prediction model will be employed. However, in the domain specific model which was proposed in Chapter 4, the evolutionary analysis and the ortholog interactions were not taken into account. Thus, in order to maintain the compatibility with the current model, incorporation of evolutionary analysis concepts is required. Even though this will add to the complexity of the model, the addition of ortholog interactions to the unified data support the comparison analysis of PPIs between several model organism which is a great advantage since

most of the laboratory work is carried out on model organism samples prior to human samples.

The ultimate goal of PPI network analysis is studying the differences between a functional and dysfunctional network. In other words, it is the comparison between healthy vs. disease states which by itself is related to the genotype-phenotype studies. Thus, a more comprehensive model which can harbour these entanglements is more beneficial as oppose to a model which focuses purely on PPIs. Therefore, the future work mainly aims at expansion of the LinkedPPI platform with regard to disease oriented pathway studies with PPI networks as its building blocks.

Bibliography

[1] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, December 2000. ISSN 1087-0156. doi: 10.1038/82360.

[2] H. N. Chua, W. K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[3] Zhang A Lin C, Jiang D. Prediction of protein function using common-neighbors in protein-protein interaction networks. *In Proceedings of the Sixth IEEE Symposium on BionInformatics and BioEngineering. 1169404: IEEE Computer Society*, page 251260, 2006.

[4] Liu K Sun Z Zhang S, Chen H. Inferring protein function by domain context similarities in protein-protein interaction networks. *BMC Bioinformatics*, 10:395, 2009.

[5] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 2008.

[6] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, 1989.

[7] Suresh Mathivanan L. Karthick K. N. Chandrika S. Sujatha Mohan Salil Sharma Stefan Pinkert Shilpa Nagaraju Balamurugan Periaswamy Goparani Mishra Kannabiran Nandakumar Beiyi She2 Nandan Deshpande Rashmi Nayak-Malabika Sarker Jef D. Boeke Giovanni Parmigiani Jrg Schultz Joel S. Bader Akhilesh Pandey T. K. B. Gandhi, Jun Zhong. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38:285–293, 2006.

[8] Paola Grandi Roland Krause Markus Boesche Martina Marzioch Christina Rau Lars Juhl Jensen Sonja Bastuck Birgit Du mpelfel1 Angela Edelmann1 Marie-Anne

Heurtier Verena Hoffman Christian Hoefert Karin Klein-Manuela Hudak Anne-Marie Michon Malgorzata Schelder Markus Schirle Marita Remor Tatjana Rudi Sean Hooper Andreas Bauer Tewis Bouwmeester Georg Casari Gerard Drewes Gitte Neubauer Jens M. Rick Bernhard Kuster Peer Bork Robert B. Russell Giulio Superti-Furga Anne-Claude Gavin, Patrick Aloy. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006.

[9] Kwang-Il I. Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, May 2007. ISSN 0027-8424. doi: 10.1073/pnas.0701361104. URL `http://dx.doi.org/10.1073/pnas.0701361104`.

[10] Iyengar R. Berger SI, Ma'ayan A. Systems pharmacology of arrhythmias. *Sci Signal*, 2010.

[11] Einat Zalckvar Adi Kimchi Martin Kupiec Eytan Ruppin Nir Yosef, Lior Ungar and Roded Sharan. Toward accurate reconstruction of functional protein networks. *Mol Syst Biol.*, 2009.

[12] Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet.*, 2007.

[13] Thierry Emonet Roger P. Alexander, Philip M. Kim and Mark B. Gerstein. Understanding modularity in molecular networks requires dynamics. *Sci Signal.*, 2014.

[14] Hao T Goldberg DS Berriz GF Zhang LV Dupuy D Walhout AJ Cusick ME Roth FP Vidal M. Han JD, Bertin N. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *nature*, 2004.

[15] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, and Tyers M. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 2006.

[16] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. DIP: the database of interacting proteins. *Nucleic acids research*, 28(1): 289–291, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.289.

[17] T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman,

Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 37(Database issue):D767–D772, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn892.

[18] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. MINT: a Molecular INTeraction database. *FEBS letters*, 513(1):135–140, February 2002. ISSN 0014-5793.

[19] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(Database issue):D525–D531, October 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp878.

[20] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1), January 2016. ISSN 1362-4962.

[21] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway Knowledgebase. *Nucleic Acids Research*, pages gkv1351+, December 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv1351.

[22] Uniprot-Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic acids research*, 37(Database issue):D169–174, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn664.

[23] Paul Flicek et al. Ensembl 2012. *Nucleic acids research*, page gkr991, 2011.

[24] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 39(Database issue): D52–7, January 2011. doi: 10.1093/nar/gkq1237.

[25] M. Ashburner, C. A. Ball, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556.

[26] You Z.H, Lei Y.K, Gui J, Huang D.S, and Zhou X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioiformatcs*, 2010.

[27] Xia J.F, You Z.H, Wu M, Wang X.L, and Zhao X.M. Improved method for predicting phi-turns in proteins using a two-stage classifier. *Protein and Peptide letters*, 2012.

[28] Lei Y.K, You Z.H, Ji Z, Zhu L, and Huang D.S. Assessing and predicting protein interactions by combining manifold embedding with multiple information integration. *BMC Bioinformatics*, 2012.

[29] Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. *Nucleic Acide Research*, 36:3025–3030, 2008.

[30] You Z.H, Lei Y.K, Gui J, Huang D.S, and Zhou X. A semi-supervised learning embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, 2010.

[31] Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jerome Wojcik, Lukasz Salwinski, et al. The HUPO PSI's molecular interaction formata community standard for the representation of protein interaction data. *Nature biotechnology*, 22(2):177–183, 2004.

[32] Kulagina NN. Friedenstein AJ, Gorskaja JF. Fibroblast precursors in normal and irradiated mouse hematopoietic organs. *Exp Hematol.*, 1976.

[33] Barry FP. Mesenchymal stem cell therapy in joint disease. *Novartis Found Symp.*, 2003.

[34] Sam M Janes Mark J D Griffiths, Dominique Bonnet. Stem cells of the alveolar epithelium. *Lancet Rev.*, 2005.

[35] Diekman BO Moutos FT Gimble JM. Guilak F, Estes BT. 2010 nicolas andry award: Multipotent adult stem cells from adipose tissue for musculoskeletal tissue engineering. *Clin Orthop Relat Res*, 2010.

[36] Dong Y. Cao C, Dong Y. Study on culture and in vitro osteogenesis of blood-derived human mesenchymal stem cells. *Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi*, 2005.

[37] Pistoia V. Uccelli A, Moretta L. Mesenchymal stem cells in health and disease. *Nat Rev Immunol.*, 2008.

[38] Soufiane Ghannam Danile Nol Farida Djouad, Carine Bouffi and Christian Jorgensen. Mesenchymal stem cells: innovative therapeutic tools for rheumatic diseases. *Nat Rev Rheumatol.*, 2009.

[39] Chullikana A Majumdar AS. Gupta PK, Das AK. Mesenchymal stem cells for cartilage repair in osteoarthritis. *Stem Cell Res Ther.*, 2012.

[40] Schroeder TM. Westendorf JJ, Kahler RA. Wnt signaling in osteoblasts and bone diseases. *Gene*, 2004.

[41] De Bari C. Augello A. The regulation of differentiation in mesenchymal stem cells. *Hum Gene Ther.*, 2010.

[42] Garrett-Beal L Yang Y. Day TF, Guo X. Wnt/beta-catenin signaling in mesenchymal progenitors controls osteoblast and chondrocyte differentiation during vertebrate skeletogenesis. *Dev Cell.*, 2005.

[43] Campanile G Cancedda R Quarto R. Martin I, Muraglia A. Fibroblast growth factor-2 supports ex vivo expansion and maintenance of osteogenic precursors from human bone marrow. *Endocrinology.*, 1997.

[44] Ishii Y Motomura H Nakamura C Ishizawa S Fujimori T Nabeshima Y Umezawa A Kanamori M Kimura T Sasahara M. Tokunaga A, Oya T. Pdgf receptor beta is a potent regulator of mesenchymal stromal cell function. *J Bone Miner Res.*, 2008.

[45] Tania Rozario and Douglas W. DeSimone. The extracellular matrix in development and morphogenesis: A dynamic view. *Dev Biol.*, 2010.

[46] X.L. Yuan, H.Y. Meng, Y.C. Wang, J. Peng, Q.Y. Guo, A.Y. Wang, and S.B. Lu. Bonecartilage interface crosstalk in osteoarthritis: potential pathways and future therapeutic strategies. *Osteoarthritis and Cartilage*, 22(8):1077 – 1089, 2014. ISSN 1063-4584. doi: http://dx.doi.org/10.1016/j.joca.2014.05.023.

[47] Larry Sherman Peter A. Herrlich. Helmut Ponta. Cd44: From adhesion molecules to signalling regulators. *Nature Reviews Molecular Cell Biology*, 2003.

[48] Jan C. Simon Ezster Tanczos Robert Peach Brett Modrell Ivan Stamenkovic Gregory Plowman Kelly L. Bennett, David G. Jackson and Alejandro Aruffo. Cd44 isoforms containing exon v3 are responsible for the presentation of heparin-binding growth factor. *J Cell Biol.*, 1995.

[49] Peter Herrlich Susanne Weg-Remers, Helmut Ponta and Harald Knig. Regulation of alternative pre-mrna splicing by the erk map-kinase pathway. *EMBO J.*, 2001.

[50] Betz BL Sciariotta J Funkhouser W Jr Murchardt C Yaniv M Sherman LS Knudsen ES Weissman BE. Reisman DN, Strobeck MW. Concomitant down-regulation of brm and brg1 in human tumor cell lines: differential effects on rb-mediated growth arrest vs cd44 expression. *Oncogene.*, 2002.

[51] Banine F Weissman BE Sherman LS Knudsen ES. Strobeck MW, DeCristofaro MF. The brg-1 subunit of the swi/snf complex regulates cd44 expression. *J Biol Chem.*, 2001.

[52] Qin Yu and Ivan Stamenkovic. Localization of matrix metalloproteinase 9 to the cell surface provides a mechanism for cd44-mediated tumor invasion. *Genes Dev.*, 1999.

[53] Caroline M. Milner and Anthony J. Day. Tsg-6: a multifunctional protein associated with inflammation. *Journal of Cell Science*, 2003.

[54] Wisniewski H.-G. Klampfer L. Oppenheim J. D. Lee, T. H. and J. Vilcek. Tsg-6: A novel secretory protein inducible by tumor necrosis factor or interleukin-1 in fibroblasts and mononuclear cells. *J Cell Biol.*, 1993.

[55] Vilcek J. Wisniewski HG. Tsg-6: an il-1/tnf-inducible protein with anti-inflammatory activity. *Cytokine Growth Factor Rev.*, 1997.

[56] Lotz M Lee S Klampfer L Lee TH Vilcek J. Wisniewski HG, Maier R. Tsg-6: a tnf-, il-1-, and lps-inducible secreted glycoprotein associated with arthritis. *J Immunol.*, 1993.

[57] Morton C. J. Parkar A. A. Hatanaka H. Inagaki F. M. Campbell I. D. Kohda, D. and A. J Day. Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. *Cell*, 1996.

[58] A. A. Parkar and A. J. Day. Overlapping sites on the link module of human tsg-6 mediate binding to hyaluronan and chrondroitin-4- sulphate. *FEBS Lett.*, 1997.

[59] Kahmann J. D. Howat S. L. T. Bayliss M. T. Parkar, A. A. and A. J. (1998). Day. Tsg-6 interacts with hyaluronan and aggrecan in a phdependent manner via a common functional element: implications for its regulation in inflamed cartilage. *FEBS Lett.*, 1998.

[60] Burgess W. H. Oppenheim J. D. Wisniewski, H. G. and J. Vilcek. Tsg-6, an arthritis-associated hyaluronan binding protein, forms a stable complex with the serum protein inter-alpha-inhibitor. *Biochemistry*, 1994.

[61] Hua J. C. Poppers D. M. Naime D. Vilcek J. Wisniewski, H. G. and B. N Cronstein. Tnf/il-1-inducible protein tsg-6 potentiates plasmin inhibition by inter-alpha-inhibitor and exerts a strong antiinflammatory effect in vivo. *J. Immunol*, 1996.

[62] Mustafa Z. Rugg M. S. Marsden B. D. Cordell M. R. Mahoney D. J. Jenkins S. C. Dowling B. Fries E. Milner C. M. et al. Nentwich, H. A. A novel allelic variant of the human tsg-6 gene encoding an amino acid difference in the cub module: chromosomal localization, frequency analysis, modeling and expression. *J. Biol. Chem.*, 2002.

[63] Hamerman D. Sandson, J. and G Schwick. Altered properties of pathological hyaluronate due to a bound inter-alpha trypsin inhibitor. *Trans. Assoc. Am. Physicians*, 1965.

[64] Mahoney D. J. Cao T. Rugg M. S. Fries E. Milner C. M. Perretti M. Getting, S. J. and A. J. (Day. The link module from human tsg- 6 inhibits neutrophil migration in a hyaluronan and inter--inhibitorindependent manner. *J. Biol. Chem*, 2002.

[65] Wisniewski H. G. Vilcek J. Maier, R. and M. Lotz. Tsg-6 expression in human articular chondrocytes. possible implications in joint inflammation and cartilage degradation. *Arthritis Rheum*, 1996.

[66] Flechtenmacher J. Buttner F. H. Karbowski A. Puhl W. Schleyerbach R. Margerie, D. and E. Bartnik. Complexity of il-1 beta induced gene expression pattern in human articular chondrocytes. *Osteoarthr. Cartilage*, 1997.

[67] Howat S. L. Dudhia J. Murphy J. M. Barry F. P. Edwards J. C. Bayliss, M. T. and A. J. (Day. Up-regulation and differential expression of the hyaluronan-binding protein tsg-6 in cartilage and synovium in rheumatoid arthritis and osteoarthritis. *Osteoarthr. Cartilage*, 2001.

[68] Dudhia J. Day A. J. Mason R. M. Flannelly, J. K. and M. T Bayliss. Tsg-6 expression is upregulated in murine str/ort model of osteoarthritis. *Trans. Ortho. Res.*, 2001.

[69] Dias A. A. M. Olszewski R. J. Klein M. J. Reis L. F. L. Mindrescu, C. and H. G. Wisniewski. Reduced susceptibility to collageninduced arthritis in dba/1j mice expressing the tsg-6 gene. *Arthritis Rheum*, 2002.

[70] Kamath R. V. Mikecz K. Brdos, T. and T. T Glant. Anti inflammatory and chondroprotective effect of tsg-6 (tumor necrosis factoralpha-stimulated gene-6) in murine models of experimental arthritis. *Am. J. Pathol.*, 2001.

[71] Kamath R. V. Brdos T. Gal I. Szanto S. Murad Y. M. Sandy J. D. Mort J. S. Roughley P. J. Glant, T. T. and K Mikecz. Cartilagespecific constitutive expression of tsg-6 protein (product of tumor necrosis factor -stimulated gene 6) provides a chondroprotective, but not antiinflammatory effect in antigen-induced arthritis. *Arthritis Rheum.*, 2002.

[72] A. Becker and J. Sandson. The source of the inter-alpha trypsin inhibitor in pathologic hyaluronate protein. *Arthritis Rheum.*, 1971.

[73] Heinegrd D. K. Wight T. N. and Hascall V. C. Proteoglycans: structure and function. *Cell Biology of the Extracellular Matrix*, 1991.

[74] Behnam Kahoussi Dominik R. Haudenschild Franqois Binette, Janet Cravens and Paul F. Goetinckg. Link protein is ubiquitously expressed in non-cartilaginous tissues where it enhances and stabilizes the interactionof proteoglycans with hyaluronic acid. *J Biol Chem.*, 1994.

[75] Tsonis PA Carlone D. Goetinck PF, Stirpe NS. The tandemly repeated sequences of cartilage link protein contain the sites for interaction with hyaluronic acid. *J Cell Biol.*, 1987.

[76] Christophe Thurieau Jean-Pierre Perin, Franqois Bonnet and Pierre JollesS. Link protein interactions with hyaluronate and proteoglycans. *The Journal OF BIOLOGICAL CHEMISTRY*, 1987.

[77] Margolis RU Margolis RK. Ripellino JA, Klinger MM. The hyaluronic acid binding region as a specific probe for the localization of hyaluronic acid in tissue sections. application to chick embryo and rat brain. *J Histochem Cytochem.*, 1985.

[78] Caterson B Heinegrd D Rodn L. Gardell S, Baker J. Link protein and a hyaluronic acid-binding region as components of aorta proteoglycan. *Biochem Biophys Res Commun.*, 1980.

[79] Goetinck PF. Stirpe NS, Dickerson KT. The chicken embryonic mesonephros synthesizes link protein, an extracellular matrix molecule usually found in cartilage. *Dev Biol.*, 1990.

[80] Roughley P Rodriguez E. Link protein can retard the degradation of hyaluronan in proteoglycan aggregates. *Osteoarthritis Cartilage*, 2006.

[81] Jr Oegema TR. Delayed formation of proteoglycan aggregate structures in human articular cartilage disease states. *Nature*, 1980.

[82] Roughley PJ Melching LI. Studies on the interaction of newly secreted proteoglycan subunits with hyaluronate in human articular cartilage. *Biochim Biophys Acta.*, 1990.

[83] Simon Tew, Peter Clegg, Christopher Brew, Colette Redmond, and Timothy Hardingham. SOX9 transduction of a human chondrocytic cell line identifies novel genes regulated in primary human chondrocytes and in osteoarthritis. *Arthritis Research & Therapy*, 9(5):R107+, October 2007. ISSN 1478-6354. doi: 10.1186/ar2311.

[84] Scherer G Yutzey KE Lincoln J, Kist R. Sox9 is required for precursor cell expansion and extracellular matrix organization during mouse heart valve development. *Developmental Biology*, pages 120–132, 2007.

[85] Victor Y. Leung, Bo Gao, Keith K. Leung, Ian G. Melhado, Sarah L. Wynn, Tiffany Y. Au, Nelson W. Dung, James Y. Lau, Angel C. Mak, Danny Chan, and Kathryn S. Cheah. SOX9 governs differentiation stage-specific gene expression in growth plate chondrocytes via direct concomitant transactivation and repression. *PLoS genetics*, 7(11), November 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002356.

[86] Czubryt MP Phan D McAnally J Qi X Shelton JM Richardson JA Bassel-Duby R Olson EN. Arnold MA, Kim Y. Mef2c transcription factor controls chondrocyte hypertrophy and bone development. *Developmental cell*, pages 377–389, 2007.

[87] Xia JF, Han K, and Huang DS. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Pept Lett*, 17:137–45, 2010.

[88] Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang. Predicting proteinprotein interactions based only on sequences information. *PNAS*, 104, 2006.

[89] Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, and Bing Wang. prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principla component analysis. *BMC Bioinformatics*, 2013.

[90] Tanford C. Contribution of hydrophobic interactios to the stability of the globular conformation of proteins. *Am.Chem.Soc.*, 1962.

[91] Krigbaum W.R an Komoriya A. Local interactions as a structure determinant for protein molecules: Ii. *Biochem Biophys Acta*, 1979.

[92] Grantham R. amino acid difference formula to help explain protein evolution. *Science*, 1974.

[93] Charton M and Charton B.I. The structural dependence of amino acid hydrophobicity parameters. 1982.

[94] Rose G.D, Geselowitz A.R, Lesser G.J, Lee R.H, and Zehfus M.H. Hydrophobicity of amino acid residues in globular proteins. *science*, 1985.

[95] Zhou P, Tian F.F, Li B, Wu S.R, and Li Z.L. Genetic algorithmbase virtual screening of combinative mode for peptide/protein. *Acta chim Sinica*, 2006.

[96] Tzu-Wen Lin, Jian-Wei Wu, and Darby Tien-Hao Chang. Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins. *Plos one*, 2013.

[97] Wassim El-Hajj Nazar ZakiEmail, Sanja Lazarova-Molnar and Piers Campbell. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics2*, 2009.

[98] B.; Huynen M.; Bork-P. Dandekar, T.; Snel. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.*, 1998.

[99] G.; Ouzounis C.; Valencia-A. Tamames, J.; Casari. A. conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, 1997.

[100] M.; D'Souza M.; Pusch-G. D.; Maltsev N. Overbeek, R. F. Use of contiguity on the chromosome to predict functional coupling. 1999.

[101] M.; Ng H.L.; Rice-D.; Yeates T.; Eisenberg D. Marcotte, E.; Pellegrini. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 1999.

[102] F.; Valencia A. Juan, D.; Pazos. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA.*, 2008.

[103] R.; Zotenko E.; Przytycka-T. Guimaraes, K.; Jothi. Predicting domain-domain interactions using a parsimony approach. *Genome Biol.*, 2006.

[104] Hui Lu Nitin Bhardwaj. Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, 2005.

[105] Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, and Masoudi-Nejad A. Locfuse: human protein-protein interaction prediction via classifier fusion using protein localization information. *Genomics*, 2014.

[106] Qiangfeng C. Zhang, Donald Petrey, José I. Garzón, Lei Deng, and Barry Honig. PrePPI: a structure-informed database of proteinprotein interactions. *Nucleic Acids Research*, 41(D1):gks1231–D833, November 2012. ISSN 1362-4962. doi: 10.1093/nar/gks1231.

[107] Lin X and Chen XW. Heterogeneous data integration by tree-augmented nave bayes for protein-protein interactions prediction. *Proteomics*, 2013.

[108] Tzu-Wen W. Lin, Jian-Wei W. Wu, and Darby Tien-Hao T. Chang. Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins. *PloS one*, 8(9), 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0075940.

[109] Chia H. Liu, Ker-Chau Li, and Shinsheng Yuan. Human proteinprotein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence. *Bioinformatics*, 29(1):92–98, January 2013. ISSN 1460-2059.

[110] Hamed Shateri S. Najafabadi and Reza Salavati. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome biology*, 9(5):R87+, May 2008. ISSN 1465-6914. doi: 10.1186/gb-2008-9-5-r87.

[111] Mahmoudian Saeideh, Yousef Abdulaziz, and Moghadam Charkari Nasrollah. Protein-protein interaction prediction using pca and svr-phcs. *Open Bioinformatics Journal*, 2015.

[112] Zhi-Ping Liu, Jiguang Wang, Yu-Qing Qiu, Ross K. K. Leung, Xiang-Sun Zhang, Stephen Kwok-Wing Tsui, and Luonan Chen. Inferring a protein interaction map of mycobacterium tuberculosis based on sequences and interologs. *BMC Bioinformatics*, 13(S-7):S6, 2012.

[113] Xinyi Liu, Bin Liu, Zhimin Huang, Ting Shi, Yingyi Chen, and Jian Zhang. SPPS: A Sequence-Based Method for Predicting Probability of Protein-Protein Interaction Partners. *PLoS ONE*, 7(1):e30938+, January 2012. doi: 10.1371/journal.pone.0030938.

[114] Mohamed Thahir P, Carbonell Jaime G, and Ganapathiraju Madhavi K. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, 11, 2010.

[115] Zhanchao Li, Lili Chen, Yanhua Lai, Yun Xie, Zong Dai, and Xiaoyong Zou. Prediction of protein phenotype based on protein interaction network by coupling genetic algorithm and k-nearest neighbor algorithm. *Anal. Methods*, 6:5281–5289, 2014. doi: 10.1039/C4AY01003E.

[116] Guo Y Vucetic S. Lan L, Djuric N. Ms-knn: protein function prediction by integrating multiple data sources. *BMC Bioinformatics.*, 2013.

[117] Mario R. Guarracino and Adriano Nebbia. Predicting protein-protein interactions with k-nearest neighbors classification algorithm. *Springer-CIBB*, page 139150, 2009.

[118] Saha Indrajit, Zubek Julian, Klingstrm Tomas, Forsberg Simon, Wikander Johan, Kierczak Marcin, Maulikb Ujjwal, and Plewczynski Dariusz. Ensemble learning prediction of proteinprotein interactions using proteins functional annotations. *Molecular BioSystems*, 2014.

[119] Zhu-Hong Yoa, Jian-Zhong Yua, Lin Zhub, Shuai Lic, and Zhen-Kun Wena. A mapreduce based parallel svm for large-scale predicting proteinprotein interactions. *Neurocomputing*, 2014.

[120] Zhu-Hong You, Lin Zhu, Chun-Hou Zheng, Hongjie Yu, Suping Deng, and Zhen Ji. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics*, 15(S-15):S9, 2014.

[121] Morihiro Hayashida, Mayumi Kamada, Jiangning Song, and Tatsuya Akutsu. Conditional random field approach to prediction of protein-protein interactions using domain information. *BMC Systems Biology*, 2011.

[122] lex Bravo Ferran Sanz Nria Queralt-Rosinach, Janet Piero and Laura I. Furlong. Disgenet-rdf: Harnessing the innovative power of the semantic web to explore the genetic basis of diseases. *Bioinformatics*, 2016.

[123] Christopher B. Barnett, Kiyoko F. Aoki-Kinoshita, and Kevin J. Naidoo. The glycome analytics platform: an integrative framework for glycobioinformatics. *Bioinformatics*, 32(19):3005–3011, 2016. doi: 10.1093/bioinformatics/btw341.

[124] René Ranzinger, Kiyoko F. Aoki-Kinoshita, Matthew P. Campbell, Shin Kawano, Thomas Lütteke, Shujiro Okuda, Daisuke Shinmachi, Toshihide Shikanai, Hiromichi Sawaki, Philip V. Toukach, Masaaki Matsubara, Issaku Yamada, and Hisashi Narimatsu. Glycordf: an ontology to standardize glycomics data in RDF. *Bioinformatics*, 31(6):919–925, 2015. doi: 10.1093/bioinformatics/btu732.

[125] Ovelleiro D Jones AR Binz PA Deutsch EW Chambers M Kallhardt M-Levander F Shofstahl J Orchard S Vizcano JA Hermjakob H Stephan C Meyer HE Eisenacher M; HUPO-PSI Group. Mayer G, Montecchi-Palazzi L. The hupo proteomics standards initiative- mass spectrometry controlled vocabulary. *Database (Oxford).*, 2013.

[126] Johan Nyström-Persson, Yoshinobu Igarashi, Maori Ito, Mizuki Morita, Noriyuki Nakatsu, Hiroshi Yamada, and Kenji Mizuguchi. Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform. *Bioinformatics*, 29(23): 3080–3086, 2013. doi: 10.1093/bioinformatics/btt531.

[127] Toshiyuki Maruyama Ikuo Kato Hiroshi Yamada Yasuo Ohno Tetsuro Urushidani Takeki Uehara, Atsushi Ono. The japanese toxicogenomics project: Application of toxicogenomics. *Mol Nutr Food Res.*, 2010.

[128] Michael Baitaluk and Julia V. Ponomarenko. Semantic integration of data on transcriptional regulation. *Bioinformatics*, 26(13):1651–1661, 2010. doi: 10.1093/bioinformatics/btq231.

[129] Jan Taubert Michael Specht Andre Skusa Alexander Regg Chris Rawlings Paul Verrier Jacob Khler, Jan Baumbach and Stephan Philippi. Graph-based analysis and visualization of experimental results with ondex. *Bioinformatics*, 2006.

[130] Torsten Blum Andreas Gerasch Michael Kaufmann Oliver Kohlbacher Jan Kntzer, Christina Backes and Hans-Peter Lenhof. Bndb the biochemical network database. *BMC Bioinformatics*, 2007.

[131] Alex Dussaq Diana Magalha es de Oliveira Emanuel Diego S. Penha, Egiebade Iriabho and Jonas S. Almeida. Isomorphic semantic mapping of variant call format (vcf2rdf). *Bioinformatics*, 16.

[132] Kokoro Ueki Gul Saad Yusuke Komiyama, Masaki Banno and Kentaro Shimizu. Automatic generation of bioinformatics tools for predicting proteinligand binding sites. *Bioinformatics*, 2016.

[133] Z. Feng G. Gilliland T.N. Bhat H. Weissig I.N. Shindyalov P.E. Bourne. H.M. Berman, J. Westbrook. The protein data bank. *Nucleic Acids Research*, 2000.

[134] Amy Ralston. Operons and prokaryotic gene regulation. *Nature Education*, 2008.

[135] Fraser P. Schoenfelder S, Clay I. The transcriptional interactome: gene expression in 3d. *Curr Opin Genet Dev.*, 2010.

[136] Sonnhammer EL. Lee JM. Genomic gene clustering analysis of pathway in eukaryotes. *Genome Res.*, 2003.

[137] Tang H. Kruglyak S. Regulation of adjacent yeast genes. *Trends Genet.*, 2000.

[138] Paola Paci Daniele Santoni, Filippo Castiglione. Identifying correlations between chromosomal proximity of genes and distance of their products in protein-protein interaction networks of yeast. *Plos One*, 2013.

[139] Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars Juhl Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(Database-Issue):D362–D368, 2017. doi: 10.1093/nar/gkw937.

[140] Hanah Margalit Einat Sprinzaka. Correlated sequence-signatures as markers of protein-protein interaction. 2001.

[141] Mirny LA. Spirin V. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA.*, 2003.

[142] Neva C. Durand Elena K. Stamenova Ivan D. Bochkov James T. Robinson Adrian L. Sanborn Ido Machol Arina D. Omer Eric S. Lander Erez Lieberman Aiden Suhas S.P. Rao, Miriam H. Huntley. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014.

[143] Chen JY. Chen H, Yu T. Semantic web meets integrative biology: a survey. *Briefing in Bioinformatics*, 2013.

[144] Marta Bleda, Joaquin Tarraga, Alejandro de Maria, Francisco Salavert, et al. CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic acids research*, 40(W1): W609–W614, 2012.

[145] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, et al. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[146] B. Alberts. The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell*, 92(3):291–294, February 1998. ISSN 00928674. doi: 10.1016/s0092-8674(00)80922-8.

[147] Andreas Ruepp et al. CORUM: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Research*, 38(Database-Issue):497–501, 2010.

[148] Andrei Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31:4157–4161, 2003.

[149] Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Ikuko N. Motoike, and Kengo Kinoshita. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, 41(D1):D1014–D1020, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1014.

[150] Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomic*, 91:243248, 2007.

[151] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, 2009. doi: 10.1126/science.1181369.

[152] Amelie Stein, Robert B Russell, and Patrick Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic acids research*, 33(suppl 1): D413–D417, 2005.

[153] Rob Jelier et al. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, 21(9):2049–2058, 2005.

[154] Maulik R Kamdar, Aftab Iqbal, Muhammad Saleem, Helena F Deus, and Stefan Decker. GenomeSnip: Fragmenting the Genomic Wheel to augment discovery in cancer research. In *Conference on Semantics in Healthcare and Life Sciences (CSHALS)*. ISCB, 2014.

[155] Ruth L Seal, Susan M Gordon, Michael J Lush, Mathew W Wright, and Elspeth A Bruford. genenames. org: the HGNC resources in 2011. *Nucl. Acids Res.*, 39(Suppl 1):D514–D519, 2011.

[156] Maulik R Kamdar, Dimitris Zeginis, Ali Hasnain, Stefan Decker, and Helena F Deus. ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. *Journal of biomedical informatics*, 47:112–130, 2014.

[157] Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Re-using cool uris: Entity reconciliation against lod hubs. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *LDOW*, volume 813 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.

[158] Steven T. Kosak and Mark Groudine. Form follows function: the genomic organization of cellular differentiation. *Genes Dev*, 18:1371–1384, 2004.

[159] IKAWA Y. MIWA Y. KIMATA K. ATSUMI, T. A chondrogenic cell line derived from a differentiating culture of at805 teratocarcinoma cells. *Cell Differentiation and Development*, 1990.

[160] Markku Tammi Jayne Lesley, Vincent C. Hascall and Robert Hyman. Hyaluronan binding by cell surface cd44. *THE JOURNAL OF BIOLOGICAL CHEMISTRY.*, 2000.