Context-specific subcellular localization prediction: Leveraging protein interaction networks and scientific texts

Lu Zhu

Bielefeld University

This dissertation is submitted for the degree of
Doctor rerum naturalium

Faculty of Technology

Disputation on September 10, 2018

Lu Zhu
Context-specific subcellular localization prediction: Leveraging protein interaction networks and scientific texts
Bioinformatics & Medical Informatics Department
German-Canadian DFG International Research Training Group (1906/1)
Faculty of Technology, Bielefeld University

Referees:
Prof. Dr. Ralf Hofestädt
Bioinformatics & Medical Informatics Department
Bielefeld University, Germany

Prof. Dr. Martin Ester
School of Computing Science
Simon Fraser University, Canada

Dr. William J Duddy
Northern Ireland Center for Stratified Medicine
Ulster University, Northern Ireland

Printed on permanent, non-aging paper according to DIN ISO 9706.

Acknowledgements

I express my deep sense of gratitude and appreciation to my advisors Prof. Dr. Ralf Hofestädt and Prof. Dr. Martin Ester for all of their support of my doctoral study and related research, for their patience, constant encouragement, and expert guidance.

Besides my advisor, I would like to thank the rest of my thesis committee for their insightful comments and encouragement, but also for the inspiring questions which incented me to widen my research from various perspectives.

My sincere thanks also go to Frank Grimm and Tobias Tekath for the stimulating discussions and the contribution to the text mining project in this dissertation. Without their precious support, it would not be possible to conduct this research task.

I would like to acknowledge funding from the international research training group GRK/1906 “Computational Methods for the Analysis of the Diversity and Dynamics of Genomes” (DiDy) for three years, and scholarship from the bioinformatics and medical informatics research group for the past months.

I am also grateful to my colleagues and friends who have supported and encouraged me along the way.

Last but not the least, I would like to thank my family: my parents and my sister for supporting me spiritually throughout writing this thesis and my life.

Abstract

One essential task in proteomics analysis is to explore the functions of proteins in conducting and regulating the activities at the subcellular level. Compartmentalization of cells allows proteins to perform their activities efficiently. A protein functions correctly only if it occurs at the right place, at the right time, and interacts with the right molecules. Therefore, the knowledge of protein subcellular localization (SCL) can provide valuable insights for understanding protein functions and related cellular mechanisms. Thus, the systematic study of the subcellular distribution of human proteins is an essential task for fully characterizing the human proteome.

The context-specific analysis is an important and challenging task in systems biology research. Proteins may perform different functions at different subcellular compartments (SCCs). Hence, the dynamic and context-specific alterations of the subcellular spatial distribution of proteins are essential in identifying cellular function. While this important feature is well-known in molecular and cell biology, most large-scale protein annotation studies to-date have ignored it.

Tissue is one particularly crucial biological context for human biology. Proteins show their tissue specificity at the subcellular level by localizing to different SCCs in different tissues. For example, glutamine synthetase localizes in mitochondria in liver cells while in the cytoplasm in brain cells. The knowledge of the tissue-specific SCLs can enrich the human protein annotation, and thus will increase our understanding of human biology.

Conventional wet-lab experiments are used to determine the SCL of proteins. Due to the expense and low-throughput of wet-lab experimental approaches, various algorithms and tools have been developed for predicting protein SCLs by integrating biological background knowledge into machine learning methods. Most of the existing approaches are designed for handling general genome-wide large-scale analysis. Thus, they cannot be used for context-specific analysis of protein SCL.

The focus of this work is to develop new methods to perform tissue-specific SCL prediction. (1) First, we developed Bayesian collective Markov Random Fields (BCMRFs) to address the general multi-SCL problem. BCMRFs integrate both protein-protein interaction network (PPIN) features and the protein sequence features, consider the spatial adjacency of

SCCs, and employ transductive learning on imbalanced SCL data sets. Our experimental results show that BCMRFs achieve higher performance in comparison with the state-of-art protein-protein interaction (PPI)-based method in SCL prediction. (2) We then integrated BCMRFs into a novel end-to-end computational approach to perform tissue-specific SCL prediction on tissue-specific PPINs. In total, 1314 proteins which SCLs were previously proven cell lines dependent were successfully localized based on nine tissue-specific PPINs. Furthermore, 549 new tissue-specific localized candidate proteins were predicted and confirmed by scientific literature. Due to the high performance of BCMRFs on known tissue-specific proteins, these are excellent candidates for further wet-lab experimental validation. (3) In addition to the proteomics data, the existing scientific literature contains an abundance of tissue-specific SCL data. To collect these data, we developed a scoring-based text mining system and extracted tissue-specific SCL associations from the abstracts of a large number of biomedical papers. The obtained data are accessible from the web based database TS-SCL DB. (4) We concluded the study with an application case study of the tissue-specific subcellular distribution of human argonaute-2 (AGO2) protein. We demonstrated how to perform tissue-specific SCL prediction on AGO2-related PPINs. Most of the resulting tissue-specific SCLs are confirmed by literature results available in TS-SCL DB.

Table of contents

List of figures **xi**

List of tables **xiii**

1 Introduction **1**
1.1 Understanding protein subcellular localizations 1
1.2 The importance of the context-specific subcellular distribution of proteins . 2
1.3 Computational prediction of protein subcellular localization 2
1.4 The aim of this work 4
1.5 Structure of this work 4

2 Background **7**
2.1 Subcellular localization 7
2.1.1 Cell and cellular compartmentalization 7
2.1.2 Protein subcellular localization 9
2.1.3 Protein translocation 10
2.1.4 Multi-localizing protein 10
2.1.5 Protein mislocalization 12
2.2 Protein-protein interaction 12
2.2.1 Types of protein-protein interactions 13
2.2.2 Databases for protein-protein interactions 13
2.2.3 Reliability of PPI data 14
2.2.4 Protein-protein interaction network 15
2.3 Basic concepts in graph theory 16
2.4 Gene co-expression network analysis 16
2.5 Bayesian inference and Gibbs sampling 17
2.6 Markov random field 18
2.7 Multi-label dataset and classification 19

2.8 Text mining data curation 20

3 Overview of protein subcellular localization prediction 25
3.1 Access to the protein SCL data 25
3.1.1 Experimental data 25
3.1.2 Knowledge-bases of protein SCLs 26
3.1.3 Limitations 26
3.2 Computational prediction method 27
3.2.1 Sequence feature based methods 27
3.2.2 Protein-protein interaction network-based approaches 28
3.2.3 Limitation of existing methods 31
3.3 Spatial adjacency of subcellular compartments 32
3.4 Direct neighbors and indirect neighbors 32
3.5 Markov random field for protein function prediction 34
3.6 From mono-SCL prediction to multi-SCL prediction 36
3.7 From generic SCL prediction to context-specific SCL prediction 37
3.8 Significance of tissue specificity in human biology 38
3.8.1 Tissue-specific SCL of proteins 39
3.8.2 Bring computational approaches to the study of tissue-specific SCL of proteins 39
3.9 Summary 40

4 Generic SCL prediction 41
4.1 The Bayesian Collective MRF Model 42
4.1.1 The weighted markov random field model 44
4.1.2 Gibbs sampler and likelihood estimation 45
4.1.3 Parameter learning 48
4.1.4 Collective MRFs 48
4.1.5 Computational complexity 49
4.1.6 Implementation 50
4.2 Experimental setup 50
4.2.1 Dataset 50
4.2.2 Evaluation 51
4.2.3 Comparison partners 52
4.3 Results 53
4.3.1 Likelihood and prediction performance 53
4.3.2 Effects of different potentials 53

4.3.3 A collective process improves the performance 54
4.3.4 Transductive learning from imbalanced MLDs 54
4.3.5 Comparison with existing methods 55
4.4 Summary 55

5 Tissue-specific SCL prediction 63
5.1 Methods 64
5.1.1 BCMRFs for predicting tissue-specific SCLs 66
5.1.2 Implementation 67
5.1.3 Data resources 67
5.1.4 Performance measures 70
5.2 Results 70
5.2.1 Statistics of the tissue-specific physical PPINs 70
5.2.2 Statistics of the tissue-specific SCLs 71
5.2.3 The impact of the noisy tissue-specific functional associations on tissue-specific SCL prediction 73
5.2.4 Genome-wide tissue-specific SCLs prediction 75
5.2.5 Predictions for novel tissue-specific protein candidate validated by text mining 76
5.3 Summary 76

6 Tissue-specific SCL Data Curation using Text mining 83
6.1 Methods 85
6.1.1 A. Retrieving relevant abstracts 85
6.1.2 B. Text preprocessing 85
6.1.3 C. Mamed entity recognition 87
6.1.4 D. Term normalization 89
6.1.5 E. Extraction and scoring of tissue-protein-SCL associations 89
6.1.6 Experimental design and evaluation 92
6.2 Results 95
6.2.1 Dictionary-based tagger 95
6.2.2 Evaluation against manual curated corpus - Tissue 96
6.2.3 Evaluation against experimental dataset - Cell lines 99
6.2.4 Creation of TS-SCL database 101
6.2.5 TS-SCL database web interface 103
6.2.6 Generality of the approach 104
6.2.7 Limitation and future direction 104

6.3 Summary 106

7 Tissue-specific subcellular distribution of the human AGO2 protein 109
7.1 Tissue-specific PPI networks of the human AGO2 protein 110
7.2 Characterization of the tissue-specific networks 111
7.2.1 Roles in RNA silencing event 111
7.2.2 Roles in mRNA splice and translation 113
7.2.3 Roles in tumorigenesis 113
7.3 Analysis of the prediction results 113
7.3.1 Generic SCLs 113
7.3.2 Tissue-specific SCLs 114
7.4 Summary 116

8 Conclusion and discussion 123
8.1 Conclusion 123
8.2 Discussion 124
8.3 Future work 126

References 129

Notations 153

List of figures

2.1 Schematic overview of the animal cell. 9
2.2 Schematic overview of intracellular protein trafficking. 11
2.3 Physical contact between two proteins. 13
2.4 Relation of types based on affinity and stability. 14
2.5 Example of simple graphs. 15
2.6 Co-expression network inference pipeline. 22
2.7 Pipeline of text mining solution. 23

3.1 Protein-protein interaction network as an undirected graph. 30
3.2 Indirect neighbors in protein-protein interaction network. 33

4.1 Binarization of multi-label MRFs. 43
4.2 Overview of the collective MRFs. 50
4.3 The overview of implementation of BCMRFs method. 57
4.4 Summarization of descriptive data from the human protein dataset. 58
4.5 Characters of PPIN dataset. 58
4.6 Relationship between the likelihood and prediction performance. 59
4.7 Performances of BCMRFs during iterations. 59
4.8 Imbalance level of each SCL class. 60
4.9 Prediction performances of four models. 61

5.1 The workflow of the tissue-specific SCL prediction based on PPINs. 65
5.2 The overview of implementation of tissue-specific BCMRFs method. 68
5.3 The property of the tissue-specific physical PPINs. 80
5.4 Comparison of protein SCLs across tissues. 81
5.5 Impact of tissue-specific functional association on performance. 82

6.1 Schematic diagram of the text mining system 86
6.2 The overview of GNormPlus method. 87

6.3 SCL mapping along the GO tree. 94
6.4 Benchmark of tissue-protein-SCL association obtained through text mining. 97
6.5 Histogram bar chart of scored true positive triplets and negative positives. . 97
6.6 Tuning the scoring parameters. 98
6.7 Comparison of the text-mined results with HPA experimentally validated cell line data. 100
6.8 Distribution of the scored triple association. 100
6.9 Illustration of web interface. 106

7.1 Best connected tissue-specific PPINs of human AGO2 protein. 118
7.2 The subcellular distribution of the interacting proteins of AGO2 across tissues. 119
7.3 The predicting subcellular distribution of human AGO2 across tissues. . . . 121

List of tables

4.1 F1 scores with/without imbalance correction. 55
4.2 F1 scores for transductive VS conventional. 55
4.3 Comparison with the method of DC-*k*NN - Multi-SCL prediction. 56
4.4 Comparison with the method of Hum-mPLoc 3.0 - Multi-SCL prediction. . 56

5.1 Mapping table from cell lines to tissues. 69
5.2 The imbalance level of SCL dataset across tissues. 72
5.3 The distribution of protein SCL across tissues. 72
5.4 tissue-specific multi-SCL prediction performance. 77

6.1 Performance of text mining system for triplet prediction 96
6.2 Accuracy of overlapped triplets. 101
6.3 Overview of TS-SCL database. 102

7.1 Interacting partners of AGO2 111
7.2 The SCL annotations of AGO2 protein. 120

Chapter 1

Introduction

1.1 Understanding protein subcellular localizations

One essential task in proteomics analysis is to explore the functions of protein in conducting and regulating the activities at the subcellular level [1]. As the eukaryotic cells and particularly the mammalian cells are highly compartmentalized, most protein activities can be assigned to particular cellular compartments. It is well known that protein functional activities highly correspond with their subcellular distribution and molecular complexing interactions [2]. A protein functions correctly only if it occurs at the right place, at the right time, and interacts with the right molecules. In other words, the functions of protein and protein interactions rely greatly on the proper localization of each protein component [3, 4]. On the other hand, the aberrant translocalization of proteins often correlates with pathological changes in cell physiology and accounts for a variety of human diseases such as Alzheimer's disease, Swyer syndrome, and various type of cancer. Hence, the mislocalization of protein makes protein translocalization a promising target for the development of therapeutic agents [5]. Therefore, the knowledge of protein SCL can provide valuable insights for understanding protein functions and related cellular mechanisms. Hence, the systematic study of protein SCLs is essential for fully characterizing the human proteome, and a major research topic in biology.

After synthesis of protein, protein can be transported into different subcellular compartments (SCCs) depending on the roles within the cell. Such translocalization of protein accomplishes the transport of material and information within and between cells. Thus, it is essential for the normal functioning of the cell. Some proteins are even transported to multiple sites simultaneously or once at the time when the protein is needed, e.g. moonlighting proteins [6] and circadian clock proteins [7]. Some of the multi-localizing proteins (MLPs) are also multi-functional proteins (MFPs), e.g. Enolase 1 fulfills different functions

in the cytosol and plasma membrane. The existence of MFPs and MLPs increases the cellular complexity because they can participate in multiple pathways or serve as regulators of transcription.

1.2 The importance of the context-specific subcellular distribution of proteins

The context-specific analysis is an important and challenging task in systems biology research, such as study on tissue-specific expression of protein of human body [8], identification of disease-specific protein-protein interaction (PPI) [9, 10], prediction of the temporal organization of cellular functions using the dynamic circadian protein-protein interaction networks (PPINs) [11], the analysis the SCL of protein under stress condition [12].

The protein function is highly dependent on the spatial distribution of many cellular components under various types of biological event, e.g. tumorigenesis [13], cellular apoptotic activity [14], and environments, e.g. stress [12], different tissues [15–17]. An example of crucial subcellular distribution is breast cancer type 1 susceptibility protein (BRCA1), well known for its nuclear, cytoplasmic trafficking in breast cancer [18]; recently, its redistribution to the cytoplasm in malignant breast cancer tissues has been supposed to be a defense mechanism of the cell probably associated with a more intense cellular apoptotic activity [19]. Moreover, glutamine synthetase (GS) is mitochondrial in liver cells and cytoplasmic in brain cells [15]. In the human tissue adrenal gland, pituitary gland and pancreas, the absence of adracalin (ALADIN) in nuclear membrane causes human triple A syndrome [16]. The dynamic alterations of subcellular spatial distribution of proteins are at least equally important to changes in total protein abundance in cellular function. However, this essential feature is long known from molecular and cell biology, but so far is often ignored in many large-scale studies. The emerging theme, which is the focus of this work, is to understand the dynamic and context specific subcellular distribution of protein.

1.3 Computational prediction of protein subcellular localization

Conventional wet-lab experiments are used to determine the SCLs of protein. The most popular wet-lab approaches such as electron microscopy, quantitative mass-spectrometric, and immunofluorescence (IF) combined with confocal microscopy are expensive and time-consuming. Unfortunately, the SCLs of a majority of proteome still remains unknown.

Owing to the automated and high-throughput nature, computational methods are appealing for the large-scale assignment of protein SCLs. Scientists have made extensive efforts to develop efficient approaches for analyzing the SCL of protein. Various algorithms and tools have been developed for predicting protein SCLs by integrating biological background knowledge into machine learning methods. Those predictions are made from the information such as protein amino acid sequence, functional domains, and motifs, protein-protein interaction (PPI), Gene Ontology (GO) annotations of protein, protein homology, key information in scientific texts, either unitarily or combined.

The existing methods have their unique strengths and disadvantages. The common limitation is that they mainly focus on the static studies of the tissue-generic subcellular distribution of the protein. Sequence-based prediction methods have been successfully used in genome-wide large-scale protein SCL annotation. However, the primary sequence of protein always reminds the same, even though the protein could have been translocated by binding to other molecules. Thus, those methods can not be applied to determine the translocalization of protein. It was shown that using the protein annotation information and protein interactions can increase the accuracy of protein SCL prediction [20, 21]. However, the existing approaches are lack of context specificity. For instance, the PPI which occurs only in brain tissue should not be used for predicting the SCL in the other tissue. The SCLs which were determined in a healthy sample might be not applicable to a study in cancer context. Using unspecified data in a context-specific study can produce unreliable results in SCL prediction. Hence, there is still room to improve in protein SCL prediction. Furthermore, the blankness of the computational approach for the systematic analysis of the protein context-specific SCL is required to be filled.

Argonaute-2 (AGO2) protein is a key player in gene-silencing pathways. It has been mostly known as a cytoplasmic protein [22]. However, more recent studies and data suggested that AGO2 is a MLP [23–26]. AGO2 is also a MFP which is involved in different biological events such as mRNAs degradation, mRNA splicing event, translation repression. Furthermore, Sharma et al. [17] demonstrated that the nuclear distribution of AGO2 occurs in a cell type- and tissue context-dependent manner. Hence, we believe that the various functions of AGO2 may correlate to its SCLs and the tissue where it expresses. A tissue-specific analysis of the subcellular distribution of AGO2 protein helps to understand its functions better.

1.4 The aim of this work

This work aims to develop efficient methods for protein tissue-specific SCLs analysis. The major components of this research are summarized in below.

- First, we developed BCMRFs to address the general multi-SCL problem. BCMRFs integrate both PPIN features and the protein features, consider the spatial adjacency of SCCs, and employ transductive learning on imbalanced SCL data sets. Our experimental results show that BCMRFs achieve higher performance in comparison with the state-of-art PPI-based method in SCL prediction.

- We then integrated BCMRFs into a novel end-to-end computational approach to perform tissue-specific SCL prediction on tissue-specific PPINs. In total, 1314 proteins which were known to localize to different SCCs in different tissues and cell lines were successfully localized. Furthermore, 549 new tissue-specific localized candidate proteins were predicted. Due to the high performance of BCMRFs on known tissue-specific proteins, these are excellent candidates for further wet-lab experimental validation.

- In addition to the proteomics data, the existing scientific literature contains an abundance of tissue-specific SCL data. To collect these data, we developed a scoring model based text mining system and extracted tissue-specific SCL associations from the abstracts of a large number of biomedical papers. The obtained data are accessible from our web-based TS-SCL database.

- We concluded the study with an application case study of the tissue-specific subcellular distribution of human AGO2 protein. We demonstrated how to perform tissue-specific SCL prediction on AGO2-related PPINs. Most of the resulting tissue-specific SCLs are confirmed by literature results available in our TS-SCL database.

1.5 Structure of this work

The dissertation is organized as follows. In Chapter 2, we begin with the fundamental knowledge which is necessary to understand this thesis. The first section introduces the cellular compartmentalization, the significance of proteins SCLs, protein translocalization event for the understanding of cellular mechanisms. One of the important tasks of this work is the prediction of protein SCLs by leveraging the PPI data using probabilistic graphical model. Thus, we explained the definition of PPI, PPIN, and major data resource and the data

reliability. Furthermore, to better understand our approaches, we recalled the basic concept of graph theory including Markov random field (MRF), Bayesian inference, Gibbs sampler, and the multi-label classification problem. In the last part of this chapter, we elucidated the data curation process using text mining technology in a nutshell.

Chapter 3 reviews the existing popular methods for experimentally detecting and computationally predicting protein SCLs. These approaches were compared from the perspectives of biological concepts, machine learning method. The limitations were also discussed including the issue of multi-label classification and the challenge of performing context-specific prediction. We described the rationale of using MRF on PPINs for protein SCL prediction. Finally, we brought out the importance of understanding the tissue-specific SCL of human proteins, and the current absence of a computational solution.

The next chapter, Chapter 4, illustrates how to improve the prediction performance of predicting multi-SCL of human MLPs in general. At first, we explained the motivation and the general idea of the approach using multi-label MRF. Next, we introduced our BCMRFs algorithm and the corresponding learning procedure. Section 4.2 details the experimental design and Section 4.3 shows the experimental results. At last, we summarized this task along with the discussion and limitations.

Chapter 5 demonstrates how to apply BCMRFs to tissue-specific PPINs for tissue-specific SCL prediction. We discussed the rationale of using tissue-specific PPINs for predicting protein tissue-specificSCL. This chapter also points out the challenge of lacking tissue-specific SCL original ('ground truth') dataset for evaluation, and provide the detailed solution. After the evaluation of the method, a large-scale analysis of tissue-specific SCL prediction for human proteome was performed.

Next, Chapter 6 presents the text mining system which is able to extract the tissue-specific SCL from scientific literature. It demonstrates a method which uses a dictionary-based approach to extract and score the triple association of tissue, protein, and SCL. The method was validated with the manually curated gold standard corpus. Thereafter, we performed a large-scale extraction against PubMed abstracts. The resulting protein tissue-specific SCL data were stored in the web-based database, TS-SCL DB, together with the experimental, knowledge-based and predicted tissue-specific SCL data.

Additionally, as important as large-scale analysis, Chapter 7 illustrates how to perform tissue-specific analysis focusing on a protein of interest. In Chapter 7, we profoundly analyzed the subcellular distribution of human AGO2 based on tissue-specific PPIs and scientific literature . The resulting tissue-specific SCLs help us to understand the specific function of AGO2 across tissues and cell types.

In the end, all the achievements and the limitations were concluded in Chapter 8. The final chapters ends by addressing further directions for the work.

Chapter 2

Background

2.1 Subcellular localization

2.1.1 Cell and cellular compartmentalization

Cells are the basic units of life that facilitate and sustain every single process within a living organism. Cells are not an unstructured mixture of proteins, lipids, ions and other molecules. Instead, the cell creates subregions, each of which allows certain cell functions to operate more effectively. As such, the subdivision of cells into discrete subcellular compartments (SCCs) enables the cell to create specialized environments for specific functions.

Despite the morphological and functional variety of cells from different tissue types, all cells share essential similarities in their compartmental organization, such as the common SCCs plasma membrane, cytoplasm, and ribosomes. The plasma membrane is a phospholipid bilayer with proteins that separates the cell from the surrounding environment and functions as a selective barrier for the import and export of materials. The plasma membrane also helps contain the cytoplasm of the cell, which provides a gel-like environment for the cell's organelles. The cytoplasm is the location for most cellular processes, including metabolism, protein folding, and internal transportation.

Unlike prokaryotic cells, eukaryotic cells (see Figure 2.1) have a nucleus enclosed within membranes. The nucleus houses the cell's genetic material DNA that determines the entire structure and function of that cell. Ribosomes are responsible for protein synthesis. Often the distinction of SCCs is made between membrane-bound and non-membrane bound organelles. The membrane-bound organelles create a physical boundary thus separating the intra-and extra-organelle space.

- Mitochondria are oval-shaped, double membrane organelles that have their own ribosomes and DNA. These organelles are often called the "energy factories" of a cell

because they are responsible for making adenosine triphosphate (ATP), the cell's primary energy-carrying molecule, by conducting cellular respiration.

- endoplasmic reticulum (ER) modifies proteins and synthesizes lipids, while the Golgi apparatus is where the sorting, tagging, packaging, and distribution of lipids and proteins takes place.
- Golgi apparatus is where the sorting, tagging, packaging, and distribution of lipids and proteins takes place. Golgi apparatus receives the entire output of de novo synthesized polypeptides from the ER and functions to posttranslationally process and sort them within vesicles destined to their proper final destination (e.g. plasma membrane, endosomes, lysosomes).
- Vesicles and vacuoles are membrane-bound organelles that function in storage and transport. Vacuoles are somewhat larger than vesicles, while the membranes of vesicles can fuse with either the plasma membrane or other membrane systems within the cell.
- Lysosomes which contains a large number of hydrolytic enzymes that are used for degrading almost any kind of cellular constituent, including entire organelles.
- Endosomes are involved in transport within the cell. They receive endocytosed cell membrane molecules and sort them for either degradation or recycling back to the cell surface. They also receive newly synthesized proteins destined for vacuolar/lysosomal.
- Peroxisomes are small, round organelles enclosed by single membranes which carry out oxidation reactions that break down fatty acids and amino acids.

In contrast, there are also non-membrane bound organelles such as the cytoskeleton and nucleoli.

- Cytoskeleton, including intermediate filaments, microfilaments, microtubules, the microtrabecular lattice, and other structures not only serve in the maintenance of cellular shape but also have roles in other cellular functions, including cellular movement, cell division, endocytosis, and movement of organelles [27].
- The most prominent substructure within the nucleus is the nucleolus which is the site of ribosomal ribonucleic acid (rRNA) transcription and processing, and of ribosome assembly.

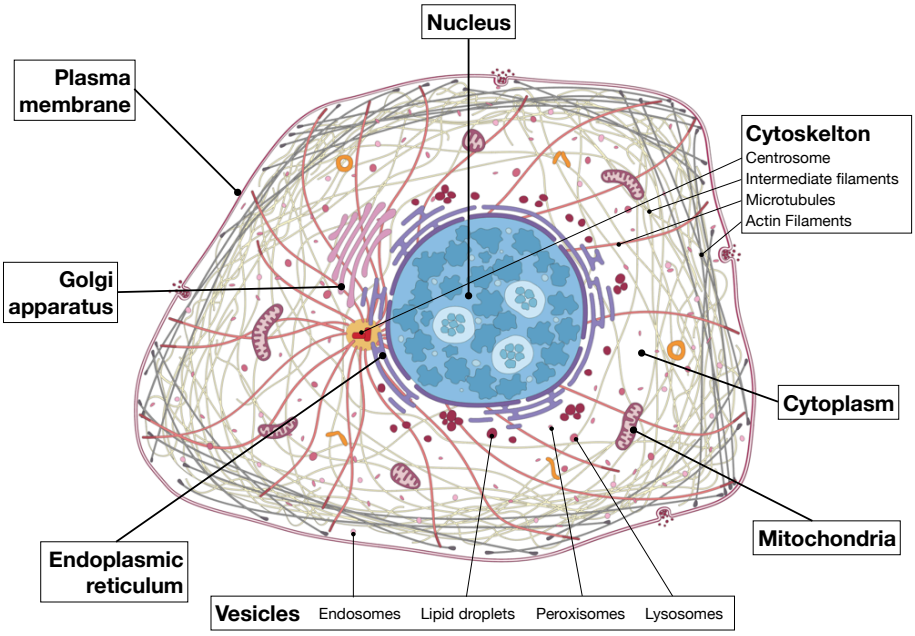


Fig. 2.1 Schematic overview of the animal cell. Eight primary SCCs, including Plasma membrane, Nucleus, Cytoplasm, Versicles, Mitochondria, Cytoskeleton, Endoplasmic reticulum, Golgi apparatus, and their substructures. Figure is adapted from Thul et al. [28].

2.1.2 Protein subcellular localization

For subcellular processes to be carried out within defined SCCs, mechanisms must exist to ensure the required protein components are present at the sites, at an adequate concentration and the correct timing. The accumulation of a protein at a given site is known as protein SCL.

One challenge in cell biology is how does the cell get materials (such as proteins, messenger ribonucleic acid (mRNA), ion) in and out across the membranes, and each compartment has its solution. The study of SCL and the transportation of the materials implicates many questions, such as: What controls the movement of a protein from one region to another? What does the protein-import material consist of? Which proteins are involved in mitochondria for instance (organelle proteome)?

The spatial partitioning of biological processes is a phenomenon fundamental to life that enables multiple processes to occur in parallel. SCLs direct the access of proteins to its interacting partners, such as other molecules and the post-translational modification machinery. Moreover, SCL is essential to protein function and its functional diversity [5].

Hence, resolving protein SCL, and the spatial distribution of the human proteome at a subcellular level can significantly increase our understanding of human biology.

2.1.3 Protein translocation

Protein translocation is a process by which proteins move between SCCs. It is a fundamental requirement for proteins to be able to exert their functions in different organelles. Short amino-acid sequences within a protein, known as signal peptides or signal sequences, can direct its localization, although translocation also occurs in the absence of these signal sequences. Protein translocation can occur co-translationally or post-translationally. Approximately half of the proteins generated by a cell have to be transported into or across at least one cellular membrane to reach their functional destination [29]. As a post-transcriptional process, some proteins translocate to the mitochondria, peroxisomes or the nucleus [30]. Whereas many proteins, including those destined for the secretory pathway and integral membrane proteins, are transported into the ER during synthesis, as the co-translational translocation [31].

Protein translocation accomplishes the movement of material and information within the eukaryotic cell and is essential for the normal activity of the cell. The protein transport machinery of cells ensures that the right amount of protein is present at the right time and place. Hung and Link [5] summarized an overview of intracellular protein trafficking and an example of protein translocation induced by peptide signal, see Figure 2.2.

2.1.4 Multi-localizing protein

Owing to the translocation, proteins which are often localized to more than one organelle, which are called multi-localizing proteins (MLPs). MLPs present several advantages for the cell, some which are crucial for cellular survival.

The multilocation of protein often happens in the following translocation scenarios. Shuttle proteins continuously switch their SCL to transport other proteins between SCCs. For instance, importin α transports protein from the cytosol to the nucleus and thus is found in both SCCs [32]. The proteins are involved in the reactions which take place in more than a single SCC, e.g. mitochondria and peroxisomes share some enzymes in their lipid metabolism [33]. Proteins translocate as a quick cellular response due to a changing environment. For example, ERBB2 protein in the plasma membrane moves to the nucleus after stimulation and change the expression pattern [34].

Furthermore, Some of the MLPs are also multi-functional proteins (MFPs). These proteins have more than one function, which might depend on the different SCLs where they are localized. These MLPs may have context-specific functions which increases the

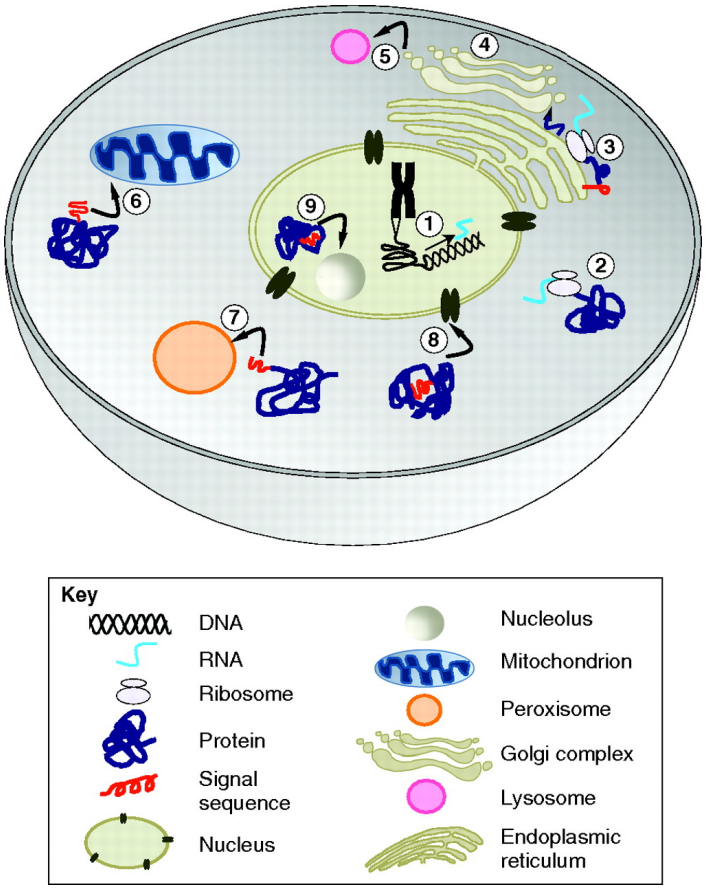


Fig. 2.2 Schematic overview of intracellular protein trafficking.The major components of the eukaryotic cell are the cytosol, the nucleus, the nucleolus, the ER, the Golgi apparatus, mitochondria and the peroxisome. Whereas gene transcription takes place within the nucleus (1), protein synthesis is confined to the cytosol and takes place either on free RNA ribosomes (2) or on ribosomes associated with the ER (3). Most proteins destined to be secreted from the cell (4), or to reside in the plasma membrane, the lysosomes (5), the Golgi apparatus or the ER, follow the secretory pathway and enter the ER before the end of translation. Proteins targeted to the mitochondria (6), peroxisome (7) and nucleus (8) are translocated after their synthesis is complete. Subnuclear localization signals include nucleolar retention signals (9), nuclear-matrix-targeting signals and signals that target proteins to splicing speckles. Figure reprinted from Hung and Link [5].

functionality of the proteome as a result. The existence of MFPs adds another dimension to the cellular complexity and offers new starting points in systems biology, because they might be involved in multiple pathways or serve as regulators of transcription. For example, a moonlighting protein alpha-enolase that acts in the cytosol as well as in plasma membrane fulfilling different functions [6].

2.1.5 Protein mislocalization

The right amount of protein presenting at the right time and place is of paramount importance for a protein to gain access to appropriate molecular interaction partners and ensures the normal operation of the cell. Abnormalities in the SCL of proteins that are important for the signaling, metabolic or structural properties of the cell can cause disorders that involve biogenesis, protein aggregation, cell metabolism or signaling [5].

Aberrantly, mislocalized proteins have been linked to human diseases as diverse as Alzheimer's disease, kidney stones, various type of cancer. The mechanisms that can lead to protein mislocalization include (i) the alterations of the protein trafficking machinery, (ii) protein targeting signals, and (iii) the changes in protein interaction or modification. More mislocalized protein associated with human disease are summarized by Hung and Link [5].

Accordingly, the cellular processes which associate events such as protein folding, cell signaling, and import and export to SCLs of proteins have been proposed as targets for therapeutic intervention. Some agents have been reported their success in influencing protein subcellular distribution in disease states. For instance, in patients with neurodegenerative diseases, affected neurons exhibit a striking redistribution of TAR DNA-binding protein (TARDBP) from the nucleus to the cytoplasm. The drug rapamycin which has been used for targeting mTOR (an essential protein kinase) can regulate and restore TARDBP SCL to the nuclear [35].

2.2 Protein-protein interaction

Protein-protein interactions (PPIs) are understood as physical contacts between proteins that occur in a cell or in a living organism in vivo. These physical contacts of high specificity are established between two or more protein molecules as a result of biochemical events steered by electrostatic forces including the hydrophobic effect (Figure 2.3). Indubitably, identification of other types of protein interactions (protein–DNA, protein–RNA, protein–cofactor, or protein–ligand) is also crucial for a comprehensive study of the interactome [36], but these types of data should not be mixed or confused with physical PPI data.

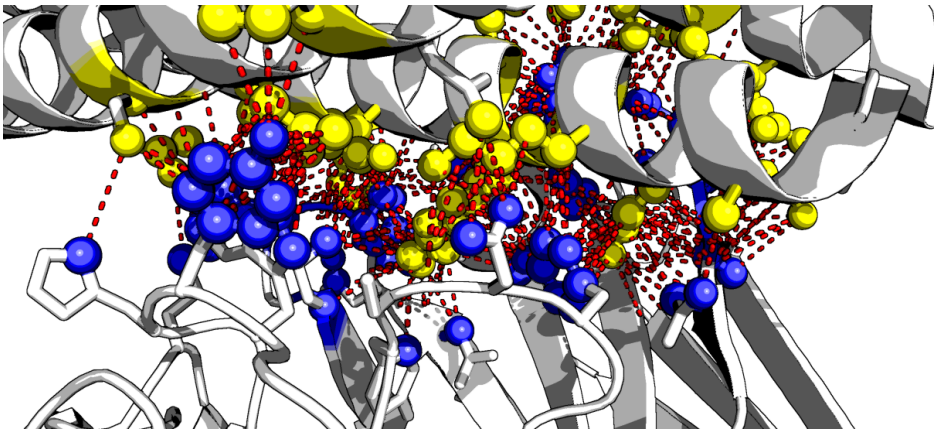


Fig. 2.3 Physical contact between two proteins. The physical PPI is the biochemical events steered by electrostatic forces (red dotted lines) between the molecules of two proteins (blue spheres and yellow spheres). Figure is reprinted from Jimwoo Leem.

2.2.1 Types of protein-protein interactions

PPIs are fundamentally characterized as stable or transient, and both types of interactions can be either strong or weak (see Figure 2.4). Stable interactions are those associated with proteins that are purified as multi-subunit complexes, and the subunits of these complexes can be identical or different [37]. Transient PPIs are expected to control the majority of cellular processes. As the name implies, transient interactions are temporary in nature and typically require a set of conditions that promote the interaction, such as phosphorylation, conformational changes or localization to discrete areas of the cell. While in contact with their binding partners, transiently interacting proteins are involved in a wide range of cellular processes, including protein modification, transport, folding, signaling, apoptosis and cell cycling [37].

2.2.2 Databases for protein-protein interactions

The repositories and databases for PPI data can be broadly classified into two types based on the content: i) Those containing interactions supported by experimental evidence, and, ii) those containing interactions derived from in silico predictions alone, or, mixed with experimentally derived PPIs. Some of the primary databases that exclusively contain experimentally derived PPI data in humans are listed here. They are Human Protein Reference Database (HPRD) [39], Reactome Knowledgebase [40], Alliance For Cellular Signaling (AfCS), DIP [41], IntAct [42], BioGRID [43],and MINT [44]. The last four are the core founders of IMEx, the international consortium of molecular interaction (MI) database providers [45]. This consortium, together with HUPO Proteomics Standards Initiative (PSI), has defined the

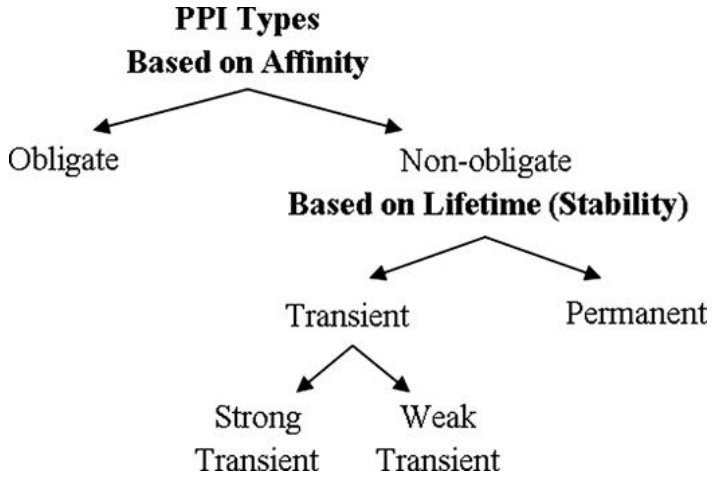


Fig. 2.4 Relation of types based on affinity and stability. Non-obligate interactions are transient but there are some examples of permanent non-obligate interactions such as enzyme-inhibitor interactions. Figure reprinted from Acuner Ozbabacan et al. [38].

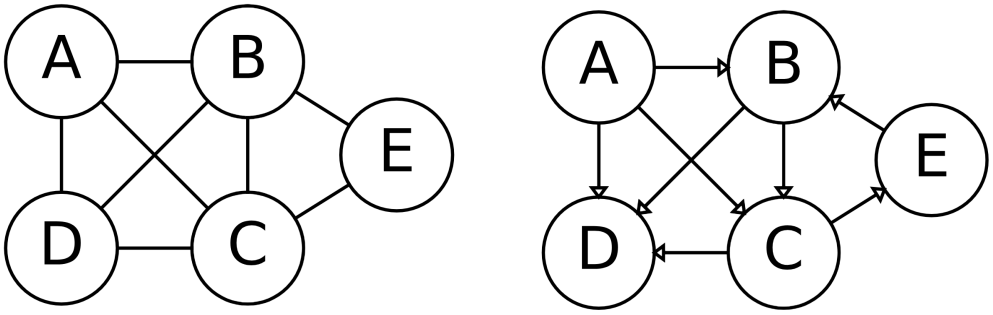
standard MIMIx (minimal information about a molecular interaction), which is proposed to improve data quality and curation of MIs [46].

2.2.3 Reliability of PPI data

Owing to technological advances, it has become increasingly feasible to detect large-scale PPI data experimentally. However, it is important to emphasize the limitations of available PPI data. Our current knowledge of the interactome is both incomplete and noisy. PPI detection methods have limitations as to how many truly physiological interactions they can detect and they all find false positives and negatives. With the accumulation of PPIs, more and more studies show the existence of a considerable amount of redundant data and false positive PPI data in the databases [47].

Because of the diversity of techniques for experimental detection, computational prediction and curation of PPI data, adequate quality assessment methods have to account for the different evidence associated with each reported interaction. An interaction of two proteins can be supported, for example, by a single concurrent mention in a scientific publication or by multiple independent experimental observations, including details such as the protein binding interface or assay parameters [48].

MINT was one of the first PPI databases to associate to each interaction a score estimating the reliability of the interaction, given the available experimental evidence [44]. The MINT score is based on a heuristic integration of the available evidence into a 'combined experimental evidence' x which is then mapped in the $[0,1]$ interval via the formula $Score = 1 - a^{-x}$. x



(a) Undirected graph **(b)** Directed graph

Fig. 2.5 Example of simple graphs.

is computed by adding up all the evidence according to the formula

$$x = \sum_i d_i e_i + \frac{n}{10} \tag{2.1}$$

where d reflects the size of the experiment. Experiments are defined on the large scale if the article reporting them reports more than 50 interactions otherwise they are defined on a small scale. This coefficient is set to 1 for small-scale and 0.5 for large-scale experiments. e depends on the type of experiment supporting the interaction and emphasizes evidence of direct interaction ($e = 1$) concerning experimental support that does not provide obvious evidence of direct interaction, i.e. Co-Immunoprecipitation (Co-IP), Pull-Down Assay, etc. ($e = 0.5$). x takes into account the number of different publications (n) supporting the interaction [49].

The MINT scoring function assigns a score close to 1 only to interactions supported by many different reports and experimental approaches while an interaction supported, for instance, by a single high throughput pull-down experiment will receive a score of 0.2.

2.2.4 Protein-protein interaction network

Protein-protein interaction networks are the networks of protein complexes formed by biochemical events and electrostatic forces that serve a distinct biological function as a complex. The protein interactome describes the full repertoire of a biological system's PPIs. In several PPI repositories, it is a straightforward process to obtain all the proteins that interact with a given query protein and from those to build a corresponding network of molecular interactions [50, 49]. Several bioinformatic tools have been developed to represent and explore such PPINs including Cytoscape [51], CELLmicrocosmos [52], VANESA [53] and many more tools are summarized in Pavlopoulos et al. [54].

2.3 Basic concepts in graph theory

Graph A simple graph G consists of a nonempty set V, called the vertices (nodes) of G, and a set E of two-element subsets of V. The members of E are called the edges (arcs) of G, and the graph can be written as $G = (V, E)$. The vertices correspond to the circles in Figure 2.5, and the edges correspond to the lines. A graph consists of vertices connected by edges. The two main categories of graphs are undirected graphs that edges do not have any particular direction (see Figure 2.5a), and directed graphs (see Figure 2.5b), where edges have direction - for example, there may be an edge from node A to node B, but no edge from node B to node A.

Graphs can be represented as a two-dimensional boolean adjacency matrix, in which the rows and columns are the sources and destination vertices, and entries in the array indicate whether an edge exists between the vertices, as in below:

$$\text{Adjacent Matrix } A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

where a non-zero value at A_{ij} indicates that node i is connected to node j.

Distance The distance between two vertices in a graph is the number of edges on the shortest path between them. In Figure 2.5, the distance from A to E is 2, whereas it is 1 from A to C.

2.4 Gene co-expression network analysis

Gene co-expression network analysis (GCNA) is a popular approach to analyze a collection of gene expression profiles. GCNA yields an assignment of genes to gene co-expression modules, a list of gene sets statistically over-represented in these modules, and a gene-to-gene network. Figure 2.6 shows the pipeline of construction of gene co-expression network. Constructing a network of genes from expression data generally consists of the following steps: 1. Prior knowledge can be used to identify guide-genes, and co-expression databases can be queried to investigate gene co-expression patterns across multiple conditions. 2. Similarity in gene expression patterns is calculated using correlation coefficients (e.g. Pearson, Spearman). A user-defined threshold (in this example set at 0.8) enables the selection of genes with high co-expression scores. Significantly co-expressed genes are reported in the binary adjacency matrix as 1. 3. A clustering algorithm is applied on the adjacency matrix to

infer networks of significantly co-expressed genes. In the resulting network, significantly co-expressed genes are depicted as numbered vertices linked by edges. A widely used approach to attach biological meaning to modules is to determine functional enrichment among the genes within a module using GCNA tools. Assuming that co-expressed gene are functionally related, enriched functions can be assigned to poorly annotated genes within the same co-expression module, and an approach commonly referred to as 'guilt by association' (GBA). GBA approaches are also widely used to identify tissue or cell type specific genes if a substantial proportion of the genes within a module are associated with a particular tissue or cell type, such as the tissue-specific interaction network database, GIANT [10].

2.5 Bayesian inference and Gibbs sampling

The Bayesian interpretation of probability is one of two broad categories of interpretations. Bayesian inference updates knowledge about unknowns, parameters, with information from data. The basis for Bayesian inference is derived from Bayes' theorem.

$$p(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{p(y)} \tag{2.2}$$

with observations y and parameter set θ. $p(y)$ will be discussed below, $p(\theta)$ is the set of prior distributions of parameter set θ before y is observed. $p(y|\theta)$ is the likelihood function, in which all variables are related in a full probability model. $p(\theta|y)$ is the joint posterior distribution of parameter set θ that expresses uncertainty about parameter set θ after taking both the prior and data into account. Since there are usually multiple parameters, θ represents a set of j parameters as $\theta = \theta_1, \ldots \theta_j$.

The goal of Bayesian inference is to maintain a full posterior probability distribution over a set of random variables. Sampling algorithms based on Monte Carlo Markov Chain (MCMC) techniques [56] are one possible way to maintain and use this distribution in the inference models. The underlying logic of MCMC sampling is the estimation of any desired expectation by ergodic averages. Any statistic of a posterior distribution can be computed as long as N samples are simulated from that distribution [57].

Gibbs sampling is one MCMC technique suitable for the task. The idea of Gibbs sampling is to generate posterior samples by sweeping through each variable (or block of variables) to sample from its conditional distribution with the remaining variables fixed to their current values. For instance, consider the random variables X_1, X_2, and X_3. We start by setting these variables to their initial values $x_1^{(0)}$, $x_2^{(0)}$ and $x_3^{(0)}$. At iteration i, we sample $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)})$, sample $x_2 \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)})$,

and sample $x_3 \sim p(X_3 = x_3 | X_1 = x_1^{(i)}, X_2 = x_2^{(i)})$. This process continues until "convergence" (the sample values have the same distribution as if they were sampled from the true posterior joint distribution). Algorithm 1 describes a generic Gibbs sampler.

Algorithm 1: Gibbs sampler

Initialize $x^{(0)} \sim x_1^{(0)}, x_2^{(0)}, \ldots, x_D^{(0)}$.

for *iteration* $i = 1, 2, \ldots$ **do**

$$
\begin{aligned}
x_1^{(i)} &\sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)}) \\
x_2^{(i)} &\sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i)}, \ldots, X_D = x_D^{(i-1)}) \\
&\vdots \\
x_D^{(i)} &\sim p(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \ldots, X_{D-1} = x_{D-1}^{(i)})
\end{aligned}
$$

end

In Algorithm 1, the posterior distribution is sampled by sweeping through all the posterior conditionals, one random variable at a time. The samples simulated based on this algorithm at early iterations may not necessarily be representative of the actual posterior distribution due to the initialization with random values. However, the theory of MCMC guarantees that the stationary distribution of the samples generated under Algorithm 1 is the target joint posterior. For this reason, MCMC algorithms are typically run for a large number of iterations (in the hope that convergence to the target posterior will be achieved). Because samples from the early iterations are not from the target posterior, it is common to discard these samples. The discarded iterations are often referred to as the "burn-in" period [58].

2.6 Markov random field

A MRF is an probabilistic graphical model that efficiently represents the joint probability distribution of a set of random variables by encoding dependencies between them. Such dependencies can be learned from data or derived from prior knowledge about the domain which is modeled. Unlike a standard classifier, an MRF enables collective inference over the entire set of known and unknown variables. MRF models have been widely used in image analysis in order to account for the local dependency of the observed pixel intensities [59]. It was also used to solve issues in system biology such as identification of differentially expressed genes [60], protein function prediction [61] involve the solution of a probability distribution defined by a discrete MRF. The concept of MRF model which is helpful for understand this thesis is briefly introduced in below. More detailed knowledge on MRF and probabilistic graphical model can be found in Koller and Friedman [62].

MRF is a undirected graph model of a joint probability distribution. It consists of an undirected graph $G = (\nu, \varepsilon)$. Consider all the nodes on graph as a set of random variables $\mathbf{X} = X_1, X_2, \ldots, X_n$, where each variable $X_i \in \mathbf{X}$ takes a value from the label set $\mathcal{L} = l_1, l_2, \ldots, l_k$. A labeling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathcal{L}^n$. The label set corresponds to segments in the case of the segmentation problem and protein function in case of the protein function prediction problem.

The corresponding Gibbs energy function E: $\mathcal{L}^n \longrightarrow \mathbb{R}$ maps any labeling $\mathbf{x} \in \mathcal{L}^n$ to a real number $E(\mathbf{x})$ called its energy. Energy function are the negative logarithm of the posterior probability distribution of the labeling. Maximizing the posterior probability equals to minimizing the energy function and leads to the maximum likelihood estimation (MLE) or maximum a posteriori (MAP) solution.

The unary potential $\phi(x_i)$ represents the cost of the assignment: $X_i = x_i$, while the pairwise potential $\phi_{ij}(x_i, x_j)$ represents that of the assignment: $X_i = x_i$ and $X_j = x_j$. Energy functions can be decomposed into sum over unary(ϕ_i) and pairwise(ϕ_{ij}) potentials as:

$$E(\mathbf{x}) = \sum_{i \in V} \phi_i(x_i) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) \tag{2.3}$$

where ν is the set of all random variables (the nodes on G) and ε is the set of all pairs of interacting variables (the edges on G). Furthermore, the potential functions could be with three or more variables [63].

2.7 Multi-label dataset and classification

In many application domains each data sample is associated with a set of labels, instead of only one class label as in traditional classification. Therefore, with Y being the total set of labels in an multi-labeled dataset (MLD) **D** and x_i a sample in dataset **D**, a multi-label classifier must produce as output a set $Z_i \subseteq Y$ with the predicted labels for the i-th sample. As each distinct label in Y could appear in Z_i, the total number of potential different combinations would be $2^{|Y|}$. Each one of these combinations is called a label set. The same label set can appear in several instances of **D**.

Imbalance of dataset In binary classification, the imbalance level is measured taking into account only two classes: the majority class and the minority class. For an imbalanced MLD, meaning that some of the labels are very frequent whereas others are quite rare, the level of imbalance of a determinate label can be measured by the imbalance ratio, IRLbl, defined in Equation (2.4). To know how imbalance is dataset **D**, the MeanIR measure is calculated

as the mean imbalance ratio among all labels, as shown in Equation (2.5). To know the significance of this last measure, the standard CV (Coefficient of Variation, Section 2.7) can be used. The SCUMBLE measure in Section 2.7 aims to quantify the imbalance variance among the labels present in each data sample [64].

$$IRLbl(y) = \frac{\arg\max_{y' \in L} \left(\sum_{i=1}^{|\mathbf{D}|} h(y', Y_i) \right)}{\sum_{i=1}^{|\mathbf{D}|} h(y, Y_i)} \tag{2.4}$$

$$MeanIR = \frac{1}{|L|} \sum_{y \in L} (IRLbl(y)) \tag{2.5}$$

$$CVIR = \frac{IRLbl\sigma}{MeanIR} \tag{2.6}$$

$$IRLbl\sigma = \sqrt{\sum_{y=Y_1}^{Y_{|Y|}} \frac{(IRLbl(y) - MeanIR)^2}{|Y| - 1}} \tag{2.7}$$

$$SCUMBLE(\mathbf{D}) = \frac{1}{|\mathbf{D}|} \sum_{i=1}^{|\mathbf{D}|} [1 - \frac{1}{IRLbl_i} (\prod_{l=1}^{|L|} IRLbl_{il})^{(1/|L|)}] \tag{2.8}$$

This characteristic makes this task even more challenging. Generally, the imbalance problem has been faced with three different approaches: data re-sampling, algorithmic adaptations, and cost-sensitive classification [65].

2.8 Text mining data curation

Text mining is also known as knowledge discovery in text data mining. It is the process of extracting previously unknown, understandable, potential and practical patterns or knowledge from a collection of massive and unstructured text data. It is a combining technique from data mining, machine learning, natural language processing, information retrieval, and knowledge management [66, 67]. Numerous text mining techniques and tools were applied in life science, e.g. mapping of genes diseases and drug discovery [68, 69], in social media, e.g. recommendation system on Facebook and Twitter [70], business intelligence, e.g. analysis of the customer satisfaction [71].

A generic overview of text mining process is illustrated in Figure 2.7 which was summarized by Rebholz-Schuhmann et al. [68]. The larger text-analytical approaches typically include:

1. Information retrieval. The tasks include data selection, document retrieval, classification, and feature extraction generally convert the documents into intermediate forms, which should be suitable for different mining purpose.

2. Information extraction from the text and are the central part of a text mining system. Information extraction comprises the identification of entities, such as genes or diseases, as well as the identification of complex relationships between those entities, including protein-protein interactions and gene-disease associations by using the algorithms including clustering, association rule discovery, trend analysis, pattern discovery and other knowledge discovery algorithms.

3. Post-processing. These tasks manipulate data or knowledge coming from information extraction step. These scientific facts can then either be used to populate databases directly or to assist the work of curation teams including the evaluation and selection of knowledge, interpretation, and visualization of knowledge. The text mining results are used to suggest hypotheses that can then be used to shape or to plan experiments to validate or to disprove the proposed hypotheses.

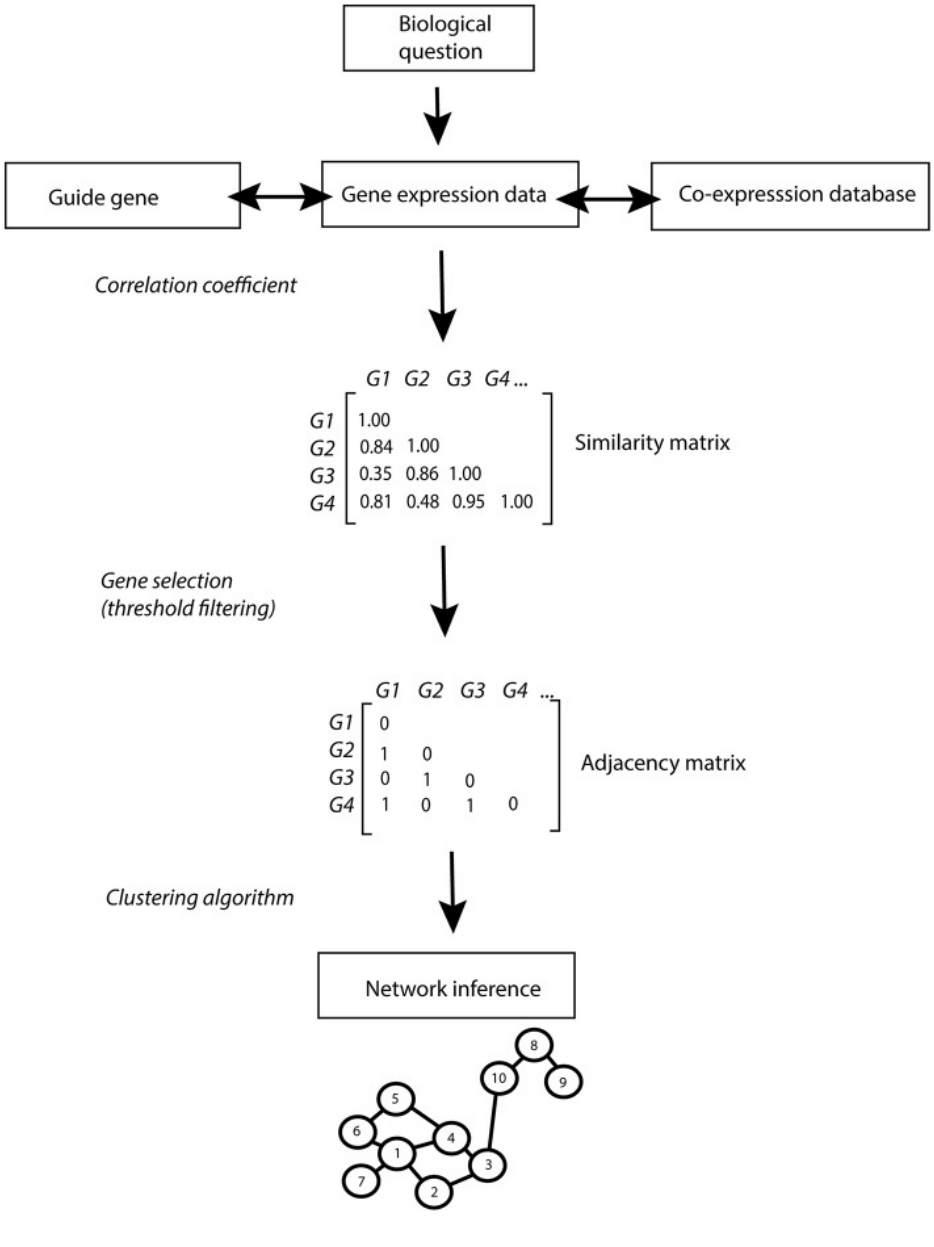


Fig. 2.6 Co-expression network inference pipeline. Figure reprinted from Serin et al. [55]

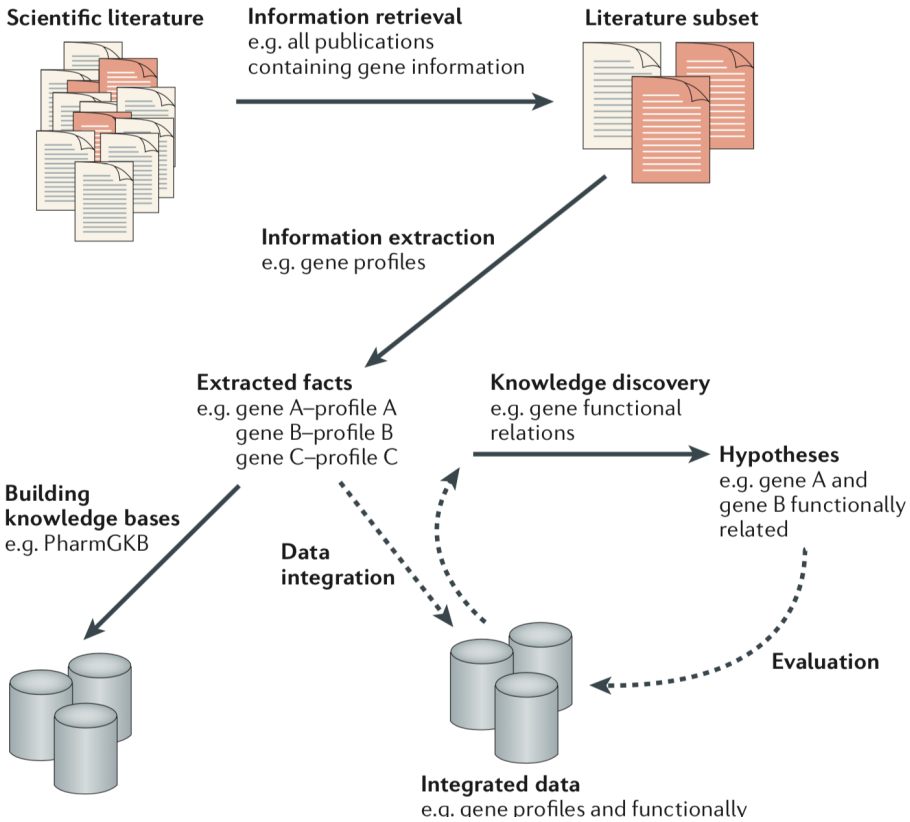


Fig. 2.7 Pipeline of text mining solution. Figure reprinted from Rebholz-Schuhmann et al. [68].

Chapter 3

Overview of protein subcellular localization prediction

Understanding of the SCLs of proteins has always been an essential aspect to discover the novel function of the protein, the primary mechanisms of the cell. We are interested in where is a protein localized in the cell? How is a MLP distributed in the cell simultaneously, or exclusively? How does the protein move from one SCC to another? Under which biological context the translocalization of protein is induced? Furthermore, how is the protein interacting with the other molecules in the cell? How to collect and access this SCL data? How to analyze the SCL data? How to interpret them along with protein functions? In this chapter, we discuss the current situation of protein SCL analysis and prediction.

3.1 Access to the protein SCL data

3.1.1 Experimental data

Conventional wet-lab experiments are used to access the SCL of proteins and also as the gold standard for validating SCL. Several wet-lab approaches for systematic analysis of protein SCLs have been developed.

- Initial research was done with specific staining and light microscopy. Closer scrutiny of micrometer- and nanometer-sized subcellular structures was later enabled by the rise of electron microscopy, which illuminated the complexity of organelles and their various positions within the cell [72].
- Quantitative mass-spectrometric readouts allow identification of proteins with similar distribution profiles across fractionation gradients [73–75] or enzyme-mediated

proximity-labeled proteins in cells [75–77]. These techniques make it possible to understand what is each component doing at the molecular level.

- Imaging-based approaches enable the exploration of the subcellular distribution of proteins in situ in a single cell and have the advantage of also effectively identifying single-cell variability and multi-organelle localization. Imaging-based approaches can be performed using tagged proteins [78] or affinity reagents [28]. Such as the immunofluorescence (IF) based approach can be combined with confocal microscopy was utilized to perform the high-resolution investigation of the spatial distribution of each protein.

Finally, genetics, in all its forms, has allowed us to dissect the structure and function of these SCLs by selective disruption of individual cell components. These experimental data can be retrieved from databases such as Human Protein Atlas (HPA) [28], LocDB [79] and ENCODE [80].

3.1.2 Knowledge-bases of protein SCLs

However, not all the experimental data are collected and accessible in databases. A vast amount of data are spread over the scientific research for various purposes. Such data require to be integrated, interpreted, standardized and enriched from literature and numerous resources to a knowledge base. UniProt Knowledgebase (UniProtKB) leads the world in providing full and comprehensive curation of the experimental data in the literature and does this in a mutually beneficial collaboration with other specialized resources. Literature-based expert curation of UniProtKB provides high-quality information for experimentally characterized proteins in a standardized and structured way using widely accepted controlled vocabularies and ontologies. Other knowledgebases of protein SCL include the Reactome Knowledgebase [40], Human Protein Reference Database [81] and Gene Ontology annotations [82].

3.1.3 Limitations

Experimentally determining the SCLs of a protein can be a laborious, expensive and time-consuming task, and manually annotating a protein, particularly identifying the massive SCL data from heterogeneous sources, is always a challenging and low-throughput task.

3.2 Computational prediction method

Over the last decades, a variety of computational methods have been developed for predicting the SCL of proteins for various organisms [83, 84], which allows us to tackle the exponentially growing number of 'omics' data and access the protein SCL data in a large scale. Predicted localization data, in particular, offer numerous insights that can assist in the prioritization of proteins for downstream analysis. Because localization and function depend on each other, a protein's localization can provide clues to its role in the cell when other information is not available.

3.2.1 Sequence feature based methods

With the rapid growth in publicly available sequence data, the computational prediction of such sequence features has become an essential aspect of biological research. By computationally identifying one or more of the signals that are known to influence protein targeting, or sequence features that correlate with a specific SCL, a protein's probable SCL can be deduced automatically using protein sequence information. These predictors utilized various methods which can be categorized into the following types:

Homology-based methods compare the SCLs of known proteins with unknown proteins. If a certain degree of similarity is found in the sequence, then it can be inferred that the unknown protein's SCL may be the same as the known protein, such as SCLpredT [85] and GOASVM [86].

Sorting-signal-based methods are more specific which recognizes the signal peptides which are responsible for protein translocation [87]. Most of the signal-based predictors aim to predict only one particular SCL, such as NucPred [88] and ChloroP [89].

Composition-based methods depend on information about the primary amino acid sequence of proteins used for information technology operations or discovery of hidden information, commonly including amino acid composition [90, 91], pseudo amino acid composition [92–94], and *n*-gram [95, 96]. Prediction results from this approach are typically less informative than those for homology- and functional domain-based methods, but in predicting SCL of unknown proteins, it is still a feasible approach.

Functional domain-based methods rely on known structures or functional data, such as protein functional domains [97–99] and motifs [100, 101], as well as information in the GO database [97, 102]. There are many learning models of research methods are used to establish the relevance of GO terms and SCL. It has been shown that GO terms can be used to advance the performance of SCL prediction. These functional data regarded as domain knowledge are highly accurate and reliable, but this approach requires manual verification of

each annotation and cannot be applied to the entirely new protein. Therefore it is usually combined with the homology-based approach [103].

Combined methods, as the name implies, these methods combine the above-mentioned protein sequence features [97, 104]. Zhou et al. incorporate multiple features, such as context vocabulary annotation-based GO terms, peptide-based functional domains, and residue-based statistical features, and use a hidden correlation modeling as a feature representation protocol which creates more compact and discriminative feature vectors by modeling the hidden correlations between annotation terms. Briesemeister et al. presented an algorithm YLoc which is based on the simple, naive Bayes classifier. They selected up to 30 most significant features from about 30 000 features from protein sequences including sorting signal, amino acid composition, and pseudo composition as well as properties such as hydrophobicity, charge, and volume of amino acids. Also, they included PROSITE motifs [105] and GO terms from close homologs. Another remarkable advantage of YLoc is that YLoc provides the details of how a prediction was made and which biological property and features of the protein was mainly responsible for it [104], whereas most of SCL prediction tools are designed as 'black box' from which the results are hard to interpret.

3.2.2 Protein-protein interaction network-based approaches

Many biological processes are mediated by dynamic interactions between proteins. Two proteins can interact with each other only if they co-occur spatially and temporally. As PPI and SCL are often discovered via separate empirical approaches, the annotations of PPI and SCL are independent and might complement each other in helping us to understand the role of individual proteins in cellular networks. We expect reliable PPI annotations to show that proteins interacting in vivo are co-localized in the same SCC.

Many studies based on high-throughput technologies have confirmed that interacting proteins tend to be localized within the same SCC, or in the physically adjacent SCCs, in various types of species. [106] reported that 76% of interactions in their yeast PPI set are localized in the same SCL, whereas a review of human PPIs based on public databases and literature curation found 52% to involve co-localized proteins plus others involving adjacent SCC [107]. These studies strengthen the assertion that a pair of interacting proteins is more likely to be co-localized in the eukaryotic cell.

The existing protein-protein interaction network (PPIN)-based methods of protein SCL prediction can be classified into three categories.

Neighborhood counting & probabilistic methods

Neighborhood counting-based approach proposed by Shin et al. [21] is the most straightforward algorithm that determines the SCL of a protein based on the known SCL of proteins lying in its neighborhood. One of the variants is the majority method, which takes the most annotated SCL terms as the prediction result. In the merged variant, for each protein, a SCL is assigned based on the union of annotations for all its interaction partners. In contrast, for the common variant method, when a protein interacts with more than one other protein only those SCL common to all its interaction partners are employed as a prediction.

Lee et al. [108] explore protein features in combination with PPINs to predict the SCLs for unknown proteins (i.e. the proteins which have no SCL information). They consider the impact of not only direct interacting neighbors but also all proteins at network distances up to five and including distance (see Section 2.3). They generated a weighted network feature vector based on each neighbor's significance and the conditional probabilities of interactions between localization pairs. Afterward, a model of SCL selection was constructed by using the supervised learning method k nearest neighbor (kNN) classifier based on both the protein feature dataset and the localization-interaction dataset to support the prediction of an unknown protein.

Du and Wang propose to use PPIN as an infrastructure integrated with the existing sequence-based predictors (Hum-mPloc 2.0 [110], Y-Loc [111]). They calculate the probability based on the neighbor protein's SCL annotation and the membership degree (see Section 2.3) of a protein in a probable SCC. In their approach, the topology of the PPIN is taken into account. Edge clustering coefficient (ECC) was firstly used for the selection of essential nodes in the context of a PPI network and was proven to be a potential indicator to whether two interacting proteins tend to have common SCC [109].

Graph theory methods

As a PPIN can be considered as a graph $G = (V,E)$ in which the nodes V represent the proteins and the edges E serve as the PPIs. The nodes V are associated with the variables X which stand for the probable SCLs label of proteins. Hence, it is natural to apply graph algorithms for the SCL assignment problem.

In contrast to the local neighborhood counting methods, the graph-based approaches are global and consider the full topology of the network. Jiang and Wu [112] applied different graph-based semi-supervised learning algorithms to assign SCL to proteins. These algorithms include χ^2-score, GenMultiCut, and Functional Flow which are initially used for protein function prediction [113]. To predict the possible SCLs of protein x, they used

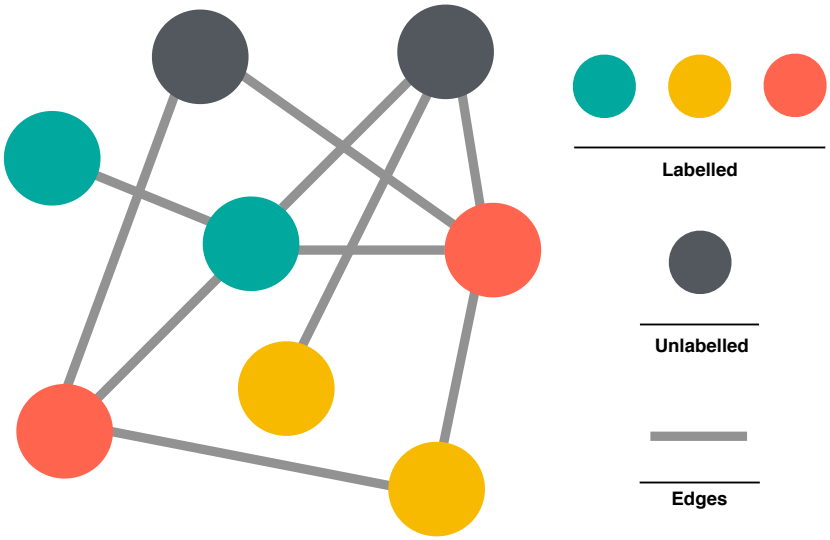


Fig. 3.1 Protein-protein interaction network as an undirected graph.The nodes represent the proteins and the edges represent the PPIs. The colors of the nodes serve as the available SCL label of proteins. Grey color indicates that the SCL information is unknown for this protein.

χ^2-score algorithm to determine the over-represented SCL from all available annotated SCL information of the neighborhood of x with a distance up to three. Whereas the GeneMultiCut algorithm [114] utilizes a cut-based methodology to maximize the number of times the same SCL annotations are associated with neighboring proteins. The task is to partition a graph in a way that each of k nodes belongs to a different subset of the partition to assign a unique SCL to all the unannotated nodes. The assignment is made by minimizing the sum of the costs (which is defined in their score function) of interacting nodes with no SCL in common.

Besides, a flow-based algorithm that simulates functional flow between proteins is applied to predict protein SCL. The proteins which the SCL information are known are treated as a 'source' of 'functional flow'. After simulating the spread of this functional flow through the neighborhoods surrounding the sources, each protein in the neighborhood is assigned with a score. This score corresponds to the amount of 'flow' that the protein has received for that function, over the course of the simulation which determines the SCL of the protein.

Methods which integrates multiple information sources

Moreover, several authors have extended the PPI data by integrating different data sources for SCL prediction. Mintz-Oron et al. [115] introduced a constraint-based method for predicting SCL of enzymes based on the embedding metabolic network, relying on a parsimony principle of a minimal number of cross-membrane metabolite transporters.

Mondal and Hu [116] integrate PPI, genetic interactions, co-expression networks. They utilize a diffusion kernel-based logistic regression (KLR) algorithm for predicting SCL using these types of protein networks. For each network, they present the data in a square matrix as well as an annotation location matrix for annotated proteins in this network. KLR model is developed based using logistic regression for predictions of un-annotated proteins.

3.2.3 Limitation of existing methods

Limitation of the sequence-based methods

Each type of sequence-based methods has pushed forward the progress on protein SCL prediction. Nevertheless, the drawbacks still exist. When the similarity between the unknown protein and the database is low, the homology-based method has a poor predictive ability. The sorting-signal-based method can only handle the proteins which carry these signal sequences. Most of the existing methods described above only predict mono-SCL proteins, those that localize to a single SCL, and they do not consider MLP. As we discussed in Section 2.1.4, identifying MLPs has high value to understanding biological functions, and there is still significant room for continued development in this area.

Stepping back to the biological point of view, the protein, to carry out different functions, can be localized in different SCCs simultaneously or at different times during various biological processes, e.g. protein translocalization (see Section 2.1.3). Although sequence-based prediction methods have been successfully applied to genome-wide large-scale protein annotations and analysis, it is hard to apply these methods to detect the translocalization of proteins because the primary sequences of the translocated protein are always about the same. The results predicted by those methods are 'static'. For the same reason, those sequence-based predictors are powerless facing the scenario of protein mislocalization and context-specific SCL.

Limitations of the PPI-based methods

The existing PPI-based prediction methods with various techniques have contributed to protein SCL assignment problem. Nevertheless, each approach has its unique advantages and drawbacks. The shortcoming of Lee et al. [117]'s approach is that each pair of proteins (direct and indirect neighbors) in the PPIs are equally treated. The χ^2 score algorithm does not consider any aspect of network topology within the local neighborhood, and cannot extend naturally to the case of weighted interaction graphs. While GenMultiCut takes more global properties of interaction maps into account, it does not reward local proximity in the graph. For example, if only two proteins have annotations in a particular network, all other

proteins will be labeled by one of these annotations, regardless of the size of the network. The metabolic network-based method is computationally more demanding and challenging to incorporate other information into their prediction model. Furthermore, an inherent limitation of metabolic network-based prediction is that it is strictly limited to metabolic enzymes. The method using diffusion KLR model is impractical for large networks because of the expensive computational cost of the required matrix exponentiation [61].

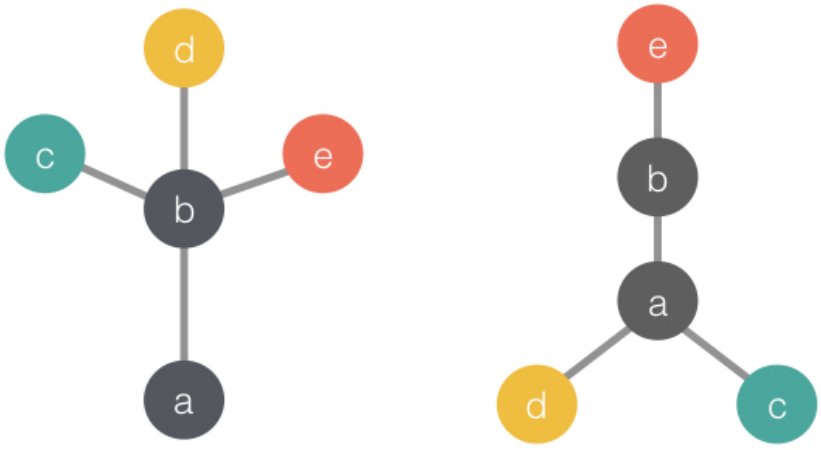
Although including various types of interactions might potentially gain more information for the SCL prediction, the expanded network would consist of a lot of unrelated proteins and interactions. The most complicated cases are protein hubs which interact widely with many other proteins without any specificity, like chaperones and ubiquitin. Mondal and Hu [116]'s result also suggests that physical PPI outperforms genetic PPI that is better than co-expression data for protein SCL prediction.

3.3 Spatial adjacency of subcellular compartments

It is worth to remind again that two physically interacting proteins must necessarily share a common SCC or an interface between physically adjacent SCCs transiently or conditionally. The SCL of a protein can, therefore, be inferred from the SCL of its interacting partners. Nonetheless, the existing PPI-based approaches only focus on the former concept (co-localizing in a common SCC). The importance of the spatial adjacency among SCCs have been underestimated. It has not been investigated whether a protein SCL (e.g. plasma membrane) can also be inferred by its interacting partners in the adjacent SCLs (e.g. Extracellular and Cytoplasm). Secondly, for the proteins whose interacting partners are poorly annotated, the information of the adjacent SCLs would be crucial as the primary resource for prediction. In this dissertation, whether the spatial adjacency among SCCs can improve PPI-based SCL prediction was discussed in Chapter 4.

3.4 Direct neighbors and indirect neighbors

Among the methods above-mentioned for protein SCL prediction, some of them consider the direct neighbors only, the others take the indirect neighbors within certain distance into account, see Figure 3.2. The concept of co-localization of the interacting neighbors in same SCC for predict SCL prediction has been well established. The indirect neighbors were often used in protein function prediction under the assumption that proteins which interact with the same proteins (i.e. distance two neighbors) may also have a higher likelihood of sharing same physics, characters and carry same functionality. However, whether the indirect neighbors



Topology A Topology B

Fig. 3.2 Indirect neighbors in protein-protein interaction network. $a-b$ is direct interacting neighbor, whereas $a-e$ is indirect neighbor. Colored nodes correspond to proteins that the SCL are known, whereas the SCL of gray nodes are unknown which is to be inferred from the SCLs neighborhood. The node a are inferred from nodes c,d,e in both graphs.

are helpful to the SCL prediction is rather debatable. The proteins which temporally interact with the same protein are not necessarily localized in the same SCC at the same time. For example, we observe that protein a interacts with protein b in ER whereas protein c interacts with protein b in nucleus. The inference of the SCL of protein c from the SCL of the indirect interacting partner protein a is incorrect.

In the neighbor counting-based approaches, the information from the indirect neighbors is helpful when the neighborhood is poorly annotated. However, this issue has less impact on the graph theory based algorithms which take global properties of the network into account including the topology and annotation information. On the contrary, when the neighborhood is well annotated, involving indirect neighbors would accumulate too many annotations, and therefore reduce the sensitivity and increase the false positive of the prediction. Furthermore, using indirect neighbors is insensitive to network topology within the local neighborhood [113]. For instance, as shown in Figure 3.2, to infer the SCL of protein a from the neighborhood with a distance of up to two, the two PPI graphs with different topologies are treated identically (a: c,d,e).

3.5 Markov random field for protein function prediction

The algorithms for protein function prediction originally inspire many of the previously discussed approaches for protein SCL prediction. These methods can be distinguished as two types. Direct annotation schemes, which infer the function of a protein based on its connections in the network, and module-assisted schemes, which first identify modules of related proteins and then annotate each module based on the known functions of its members. When using the only PPIN data, the direct methods were slightly superior to module-assisted ones, with MRF and Markov clustering (MCL) being the leading techniques for direct and module-assisted function prediction, respectively [118].

Deng et al. [119] pioneer to use MRF for protein function prediction based on PPIN. Later, the parameter estimation task of this method was improved by Kourmpetis et al. [61]. The MRF model provides a probabilistic framework for simulating the mutual influence of random variables via a neighborhood system. Given a network of influence, the state of any random variable is assumed to be independent of all other random variable states given those of its immediate neighbors. In the function prediction setting, each random variable corresponds to a protein, and its states correspond to certain functional annotations. The joint distribution of the random variables can be shown to factorize over the cliques of the network. Therefore, the probability of a certain assignment of discrete states $X = (x_1, \ldots, x_N)$ can be written as in below.

$$p(x) = \frac{1}{Z}\exp(-H(x)) = \frac{1}{Z}\exp\left\{-\sum_{c\in C} H_c(x_c)\right\} \tag{3.1}$$

where N is the total number of variables, Z is a normalizing constant, C is the set of all cliques in the network, H_c is a potential function associated with clique c, and x_c is the assignment of states to the members of c. Inference in this general model is computationally intractable. Hence it is common to assign 0 potentials to all cliques of size greater than 2, and further homogenize the model by associating the same potential function with all cliques of the same size. For such a homogeneous second-order MRF, we have the following equation.

$$H(x) = \sum_{v\in V} H_1(x_v) + \sum_{(u,v)\in E} H_2(x_{(u,v)}) \tag{3.2}$$

To obtain a second-order MRF model, they assumed that the probability of the binary annotation $[0,1]$ over the entire network is proportional to $e^{\alpha N_{01}+\beta N_{11}+N_{00}}$, where α, β are parameters for weighting the contributions of the different terms and N_{ij} is the number of interacting pairs with assignment i, j (unordered). Combining the *a priori* probability of an

assignment with N_1, which depends on the frequency f of the function and N_1 is proportional to $(\frac{f}{1-f})^{N_1}$. A homogeneous second-order MRF is therefore

$$\begin{aligned} H(x) = &- \log \frac{f}{1-f} \sum_{v \in V} x_v - \beta \sum_{(u,v) \in E} x_u x_v \\ &- \alpha \sum_{(u,v) \in E} [x_u(1-x_v) + (1-x_u)x_v] \\ &- \sum_{(u,v) \in E} (1-x_u)(1-x_v) \end{aligned} \tag{3.3}$$

Hence, the probability that protein v is assigned with the function given the annotations of its neighbors $N(v)$ is

$$\begin{aligned} P(x_v = 1 | x_{N(v)}) = logit(&\log \frac{f}{1-f} + \beta N(v,1) \\ &+ \alpha(N(v,1) - N(v,0)) - N(v,0)) \end{aligned} \tag{3.4}$$

where $N(v,i)$ is the number of neighbors of v that are assigned with $i \in \{0,1\}$ and logit is the logistic function $logit(x) = 1/(1+e^x)$. [119] estimate the two parameters of the model using a quasi-likelihood method [120] and apply Gibbs sampling (see Section 2.5) to infer the unknown functional annotations. In the field of protein function prediction, the MRF showed its superior owing to the use of a more sophisticated probabilistic model [118].

Rationale of using MRF for protein SCL prediction

MRF had success in solving various problems, such as image segmentation, image restoration in computer vision [121], the identification of differentially expressed genes in systems biology [60], and the protein function prediction which was discussed above. Over and above than that, MRF is especially suitable for protein SCL prediction due to the following reasons:

- Interacting proteins share a common SCC or physically adjacent SCCs. Therefore, the biological reasoning and the dependency of SCLs between interacting proteins are more substantial than their functions.
- Potentially better precision due to the fewer annotation terms in comparison with protein function categories.
- The protein SCL prediction is a much broader classification issue than the protein function, which means that the network topology may provide sufficient evidence for its inference using neighborhood-based approach.

3.6 From mono-SCL prediction to multi-SCL prediction

In Section 2.1.4, we show the significance of MLPs. Efficient prediction of multi-localization (ML) of proteins has always been a challenging task in protein SCL prediction which is a multi-label classification (MLC) problem.

There are two main approaches to tackle the MLC problem: Data transformation and algorithm adaptation. The former approach aims to produce from a MLD to another dataset or group of datasets that can be processed with traditional classifiers, while the objective of the latter is to adapt existent classification algorithms to work with MLDs. Among the transformation methods, the most popular are those based on the binarization of the MLD. It includes the binary relevance (BR) [122], also known as 'one vs. all', the pairwise comparison [123], and the label powerset (LP) transformation [124], which produces a multi-class dataset from an MLD considering each label set as one class. In the algorithm adaptation approach there are proposals of algorithms based on nearest neighbors such as multi-label *k*NNs [125], multi-label neural networks [126], multi-label decision tree [127], and multi-label support vector machines (SVMs) [128].

Recently, several multi-label classification methods have been employed for SCL prediction in different species, such as methods using multi-label ensemble classifier [92], multi-label *k*NNs [129] multi-label SVMs [130], with feature construction of protein sequences, such as *n*-gram, Chou's PseAAC representation, and GO.

However, as the most of MLD, the datasets containing MLPs are typically heavily imbalanced (see Section 2.7). The learning from an imbalanced multi-label classification is a well-known challenge in data mining [64]. The imbalance issue occurred in protein SCL problems have not been profoundly addressed.

Learning from imbalanced multi-labeled dataset

The learning from imbalanced data problem is founded on the different distributions of class labels in the data, and it has been thoroughly studied in traditional classification. Generally, the imbalance problem in MLD has been faced with two different approaches:

- Resampling algorithms. It consists of the label powerset based resampling algorithms, the individual label evaluation resampling algorithm [64], and the inverse random undersampling proposed by [131]. These algorithms rely on the rebalancing of class distributions through either deleting instances of the most frequent class (undersampling) or adding new instances of the least frequent one (oversampling). The advantage of this approach is that it can be applied as a general method to solve the imbalance

problem, independently of the classification algorithms used once the datasets have been preprocessed.

- Algorithmic adaptations. Most of the published algorithms aim to deal with the imbalance problem through the algorithmic adaptation of their MLC classifiers or the use of ensembles of classifiers. Those approaches are classifier-dependent, instead of general application methods able to work with other MLC learning algorithms. Ensemble multi-label learning is a method based on the use of various algorithms to build an ensemble of MLC classifiers. It exits two problems simultaneously, learning from imbalanced data and capturing correlation information among labels. He et al. [132] proposed an algorithm which is based on the use of Gaussian Process, a Bayesian method used to build non-parametric probabilistic models. Utilizing a covariance matrix the correlations among labels are obtained, and the imbalance is fixed to associate a weight coefficient to each sample.

3.7 From generic SCL prediction to context-specific SCL prediction

This section discusses another critical problem in protein SCL prediction which has so far received little attention: how to annotate protein SCL in a context-specific manner? The context means the research background of the study which could be disease, tissue, culture environment, stage of cancer and so on. An increasing number of examples indicate that in higher organisms, functional plasticity may be the rule rather than the exception [6, 133]. A protein may localize in different SCCs, which acquires different functions under different endogenous or exogenous conditions. Pinto et al. [134] showed that dynamic redistribution of multitudinous proteins to different SCCs in response to cellular functional state is a crucial characteristic of cellular function that seems to be at least as important as overall changes in protein abundance [134].

However, current SCL databases, such as UniProtKB, all annotate protein SCLs without specifying the necessary context. Notably, all of these previously discussed methods have difficulty predicting the context-specific or dynamic behavior of protein SCL. The main difficulty in predicting such dynamics is the lack of known protein SCL and functions under the specific condition(s), which are required for generating a prediction model in the training stage (ground truth dataset).

One possible solution is to find dynamic network modules in gene expression networks constructed under specific conditions. In the field of protein function prediction, Wallach et al.

[11] constructed a dynamic circadian PPI network predicting the PPI timing using circadian expression data. They predict that circadian PPIs dynamically connect many critical cellular processes (signal transduction, cell cycle, etc.) contributing to the temporal organization of cellular physiology in an unprecedented manner. Lee et al. [117] proposed an integrative computational framework for mapping stress-induced localization and mislocalization of proteins on a proteome-wide scale. They mapped the locations of over 10,000 proteins in the healthy human brain and glioma, out of which over 150 have a substantial likelihood of mislocalization under glioma. Fifteen of these mislocalizations have been confirmed. The most common type of mislocalization occurs between the endoplasmic reticulum and the nucleus [117]. A surprising number of proteins translocate from the mitochondria to the nucleus or from endoplasmic reticulum to Golgi apparatus under stress [12]. Later, similar research was carried out by Liu and Hu [135]. They developed an approach for discovering mislocalization related disease/cancer genes based on aberrant gene expression data (co-expression data) and diffusion kernel-based logistic regression for SCL prediction. Their approach has identified several cancer genes reported by genomic study, through which cancer may be related to their mislocalization within the cell [135].

3.8 Significance of tissue specificity in human biology

Although all human cells carry out common processes that are essential for survival, in the physical context of their tissue environment, they also exhibit unique functions that help define their phenotype. The tissue-specific genes with elevated expression in a particular tissue are interesting as a starting point to understand the biology and the function of this part of the human body [8], whereas housekeeping genes are involved in basic cell maintenance and, therefore, are expected to maintain constant expression levels globally [136].

Protein molecules constitute the primary building blocks of cells and mediate most cellular processes. In human, they are encoded by over 22,000 different genes, which give rise to many more proteins through alternative splicing mechanisms. These numerous proteins do not work in isolation: instead, they interact with each other and with other types of molecules to form complex cellular machines and to pass signals within cells and across tissues.

While knowledge of context-specific PPIs is limited, we are witnessing a rapid accumulation of context-specific molecular expression profiles. Many studies revert to identifying PPIs that are feasible in these contexts with the assumption that a PPI is possible within a specific context if the corresponding proteins are expressed in that context, especially

the tissue-specific PPIs [10, 137]. The tissue-specific PPIs of human protein have been intensively studied and well explored [9].

3.8.1 Tissue-specific SCL of proteins

The lack of context-specificity also exists in protein SCL annotation data that were measured in different tissues and cell types. The human protein atlas (HPA) has reported that more than 50% of the analyzed proteins in their database were identified to localize to more than one compartment at the same time. These MLPs may have context-specific functions increasing the functionality of the proteome [28]. Furthermore, from some proteins, we do observe the tissue specificity at the subcellular level. First of all, some particular SCCs exist only in specific tissues. For instance, the sarcolemma is a unique SCC in muscle tissue. Moreover, the spatial distribution of the SCLs of the protein in a cell could be different from one tissue to another, which depends on the functions of the protein in the specific tissue. For example, glutamine synthetase (GS) is mitochondrial in liver cells and cytoplasmic in brain cells [15]. In the human tissue adrenal gland, pituitary gland and pancreas, the absence of adracalin (ALADIN) in nuclear membrane causes human triple A syndrome [16]. Therefore, understanding the specific SCLs of human protein in different tissues and organs of the human body would significantly increase our knowledge of human biology.

3.8.2 Bring computational approaches to the study of tissue-specific SCL of proteins

Despite the growing understanding of the tissue-specific proteome, to the best of our knowledge, there is not yet a computational method for predicting tissue-specific SCL. The success of the studies in protein mislocalization [135, 117] and the protein circadian-specific function prediction [11], which are based the disease-specific and circadian PPINs respectively, indicate the potentials to use dynamic tissue-specific PPIN to solve the tissue-specific SCL prediction problem.

Given the lack of tissue-specific PPIs that were measured in different tissues and cell types, many studies revert to identifying PPIs that are possible in that tissue. Their underlying assumption is that a PPI is feasible within a specific tissue if the corresponding proteins are expressed in that tissue. Additionally, co-expression has often been based on RNA levels, as protein expression levels were rarely available. This approach had been used extensively for analyzing tissue interactomes [138, 10, 139]. A study shows that PPIN that appear to be tissue-specific or global expressed have distinct topological features relative to the generic human interactome or each other [9]. Bossi and Lehner [140] found extensive

direct interactions between globally expressed proteins and tissue-specific proteins and demonstrated the evolution of tissue-specific functions through the modification of core cellular processes [140]. Regarding how to realize the prediction tissue-specific SCL of human proteins based on dynamic tissue-specific PPINs, the experiments and evaluations can be found in Chapter 5 of this dissertation.

3.9 Summary

This chapter rolled out a profound analysis of the related works of the analysis and prediction of protein SCLs.

In the general protein SCL prediction using computational approach, the protein sequence, the physicalPPI data two crucial data resource to protein SCL prediction. However, there is room to improve the performance of the general protein SCL prediction. Taking the spatial adjacency of SCCs into account, using a more sophisticated machine learning method on the high imbalance MLD seems promising starting point.

The knowledge of the tissue-specific SCLs can enrich the human protein annotation, and thus will increase our understanding of human biology. The blankness of computational approach to performing tissue-specific SCL prediction should be filled. Exploration of the tissue-specific PPINs is one of the potential solutions of the tissue-specific SCL prediction problem.

Therefore, the two major goals of this work are to improve the protein SCLs prediction and to develop methods for performing tissue-specific SCL prediction and analysis.

Chapter 4

Generic SCL prediction

Detailed molecular knowledge of the human proteome has become an important asset in the understanding of human biology and disease. Rapid advances in biotechnology have made available a variety of high-throughput experimentally obtained proteomics and interactomics datasets [39, 141], and knowledge of SCL of proteins can provide important insights for understanding their functions in cells and the mechanism of disease [5]. Owing to the annotation efforts of model organism databases, high-quality subcellular localization information for human proteins can be obtained from various curated sources. However, manually annotating a protein, especially determining the subcellular localization using the enormous data from heterogeneous source, is always a challenging and low-throughput task. A variety of computational methods have been developed for predicting the SCL of proteins for various organisms [83, 84] in the past decade. Nevertheless, there are relatively few efficient specific prediction tools for human proteins in the face of rapidly increasing numbers of newly identified proteins.

Protein features, especially the sequence-based features, are always the essential part in various protein SCL predictors [142, 111, 92]. To carry out different functions, one protein can be located in different SCCs simultaneously or at different times during different biological processes, e.g. protein trafficking. Sequence-based prediction methods have been successfully applied to genome-wide large-scale protein annotations and analysis. However it is hard to apply these methods to detect the translocalization of proteins due to the fact that the primary sequences of the translocated protein are always about the same. The biological functions of proteins are carried out by interacting with other proteins. To interact, proteins (or any other molecules) must necessarily share a common SCC, or an interface between physically adjacent SCCs, transiently or conditionally. The SCL of a protein can therefore be inferred from the SCL of its interacting partners. Hence biological network information can complement feature-based approaches to SCL prediction.

Several methods have been developed which take advantage of PPINs to predict the SCL of proteins for different organisms using data integration from multiple data sources [112, 21, 108]. However, these approaches mainly focus on the co-localization (in the same SCL) of interacting proteins. The importance of the spatial adjacency among SCCs was underestimated. It was not investigated whether a protein SCL (e.g. Plasma membrane) can be also inferred by its interacting partners in the adjacent SCLs (e.g. Extracellular and Cytoplasm). Secondly, for the proteins whose interacting partners are poorly annotated, the information of the adjacent SCLs can be used as the major prediction resource. In this chapter, whether the spatial adjacency among SCLs can improve PPI-based SCL prediction performance was investigated.

Conventional machine-learning approaches, such as supervised learning, predict protein SCLs by extracting information only from existing annotation. However, the number of unreviewed proteins increases at a remarkably faster rate than that of experimentally-annotated ones. It was shown that transductive leaning approaches are able to take advantage of the large number of available unknown data to improve the accuracy of classification [143, 144]. On the other hand, proteins are often annotated with multiple SCLs. The MLDs of protein SCLs are typically heavily imbalanced. The learning from an imbalanced multi-label classification is a well-known challenge in classification [64].

A MRF is a graphical model of a joint probability distribution. Many problems in computer vision such as image segmentation, image restoration and systems biology such as identification of differentially expressed genes [60], protein function prediction [61] involve the solution of a probability distribution defined by a discrete MRF. This chapter describes the algorithm BCMRFs for predicting the multi-SCLs of human proteins considering features of PPINs, the proteins features, the spatial adjacency of SCCs and the imbalance of the dataset.

The rest of the chapter is organized as follows: Section 4.1 introduces the MRFs and the corresponding learning procedure. Section 4.2 details the experiment protocol and Section 4.3 shows the experimental results. The conclusion this task along with the discussion of directions of future work is in the end of the chapter.

4.1 The Bayesian Collective MRF Model

The basic definitions and notations of MRF can be found in Section 2.6. The rationale of using MRFs for protein SCL prediction and the BCMRFs for predicting the multi-SCLs of human proteins are proposed in below.

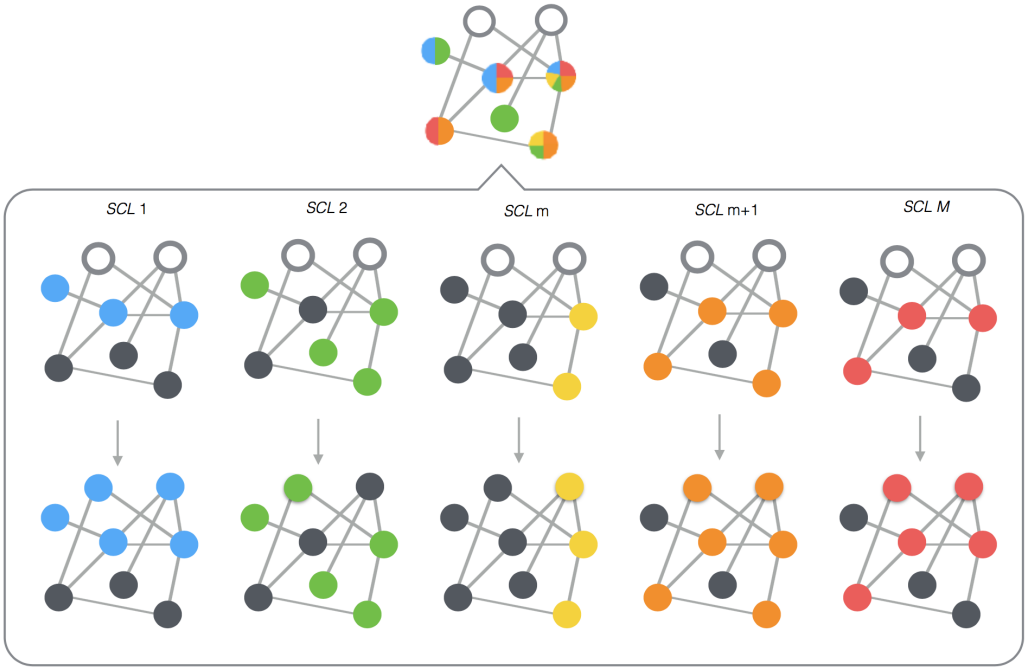


Fig. 4.1 Binarization of multi-label MRFs. The graph with multi-colored nodes on the top represents the general PPIN. Each node on the graph represents a protein associated with in total M SCL annotation terms. This network can be derived to M PPI networks for each single SCL term. The nodes are colored or in grey which represent 1 and 0 respectively if the SCL annotation of the protein is available for any of the M SCL terms. Otherwise, the node is not colored. The nodes (proteins) are in need to be assigned with SCLs.

As previously discussed above and in Chapter 3, the SCLs of a protein can be inferred by the SCLs of its physically interacting proteins. A physical PPIN $G = (P,I)$ with N proteins, $N = |P|$, that are assigned in M different SCLs in total fullfills the definition and properties of a MRF. It's reasonable to apply MRFs on PPIN to predict the SCLs of a set of proteins in the network. Moreover, a PPIN in which each protein is labeled by single SCL or multi SCLs can be considered as multi-label MRFs. Inspired by the statistical power of MRF models, MRFs were applied on PPIN for solving protein SCL prediction problem.

Using the binary reference approach [122], for the SCL noted as $l_m, 1 \leq m \leq M$, the network is encoded in an N-dimensional vector $\mathbf{x} = \{x_1, \ldots, x_N\}$, where $x_i = 1$ if the protein $p_i, 1 \leq i \leq |P|$ is assigned with l_m, else $x_i = 0$. The multi-label classification problem is thus reduced to multiple binary classification problems (Figure 4.1). For each SCL, a corresponding binary MRFs was built to predict SCL labeling of unknown proteins by maximizing the posterior probability distribution of the SCL labeling of proteins. The following elements are used in the MRFs model: 1. prior probability of any protein being

located in l_m, 2. the number of interacting neighbors being located in l_m, 3. the number of interacting neighbors being located in the adjacent SCLs of l_m, and 4. the sequence-based features of protein.

Meanwhile, the quality of PPI data and the connectivity of PPIN is crucial for inferring the SCL of a protein by its interacting neighbors. However the confidence of PPIs varies from one to another depending on the method, and the size of experiment etc. [49]. To balance of having a high quality of PPIN and reduce the risk of losing valuable information by removing too many edges, the confidence scores of the PPIs was employed to weighted MRFs. The detailed method is described in the following sections.

4.1.1 The weighted markov random field model

By definition (see Section 2.6), the posterior distribution $Pr(\mathbf{x})$ over the SCL labelings of the MRF is a *Gibbs* distribution which can be written in the form:

$$Pr(\mathbf{x}) = \frac{1}{Z}\exp(-E(\mathbf{x})) \tag{4.1}$$

where Z is a normalizing constant known as the partition function. $E(\mathbf{x})$ is the energy function of the MRFs which is defined as follows:

$$\begin{aligned} E(\mathbf{x}) = -(&\sum_{i\in\nu}\phi_i^S(x_i) + \sum_{i\in\nu}\phi_i^F(x_i,F_i) + \sum_{i,j\in\varepsilon}\omega_{i,j}\phi_{ij}^P(x_i,x_j) \\ &+ \sum_{i,j\in\varepsilon}\omega_{i,j}\phi_{ij}^A(x_i,x_j,A_{ij})) \end{aligned} \tag{4.2}$$

with the unary potential

$$\phi_i^S = \begin{cases} 0 & x_i = 0 \\ \alpha & x_i = 1 \end{cases} \tag{4.3}$$

where α is the probability of a protein located in l_m. $\phi_i^F(x_i,F_i)$ is feature-based potential. F_i is a vector that includes the features for protein i. Conditional probability of a protein p_i being located in l_m given its features $Pr(x_i = 1|F_i)$.

$$\phi_i^F(x_i,F_i) = \begin{cases} 0 & x_i = 0 \\ \eta Pr(x_i = 1|F_i) & x_i = 1 \end{cases} \tag{4.4}$$

with

$$Pr(x_i = 1|F_i) = Pr(x_i = 1)\prod_{f=1}^{F} Pr(F_i^f|x_i = 1) \tag{4.5}$$

It includes thirty features which are generated from previous widely used sequence-based protein SCL predictor YLoc [111] into the BCMRFs models. These features include various types from simple amino acid composition to annotation information. Certain features are general such as protein size, number of small residues etc., while others specifically describing one certain SCL only. η is an unknown parameter associate to the ensemble of the 30 features F_i for protein i. The class priors and the feature probability distributions are estimated using the entropy-based supervised discretization of the training data. The final probabilities are obtained by normalizing the posterior such that the sum of all posterior is one. η together with other unknown parameters are estimated during parameters learning process.
ϕ^P is the pairwise potential of two proteins locating in l_m.

$$\phi_{ij}^P(x_i,x_j) = \begin{cases} 0 & (i,j) \notin \varepsilon \\ 0 & (i,j) \in \varepsilon \ \& \ x_i = x_j = 0 \\ \beta^{11} & (i,j) \in \varepsilon \ \& \ x_i = x_j = 1 \\ \beta^{10} & (i,j) \in \varepsilon \ \& \ x_i = 1 - x_j \end{cases} \tag{4.6}$$

where $\omega_{i,j}$ is a constant parameter, the confidential score of the interaction between P_i and P_j. $\phi_{ij}^A(x_i,x_j,A_{ij})$ is the potential which depends on if the protein p_i interacts with the proteins locating in the adjacent SCLs of l_m,

$$\phi_{ij}^A(x_i,x_j,A_{ij}) = \begin{cases} 0 & i,j \notin \varepsilon \\ \sum_{h=1}^{H} \mu_h A_{ij}^h & (i,j) \in \varepsilon \ \& \ x_i = 1 \end{cases} \tag{4.7}$$

where H is the total number of adjacent SCLs of SCL l_m. Given a set of H adjacent SCLs of SCL l_m, for each protein p_i which has N_{ne} of neighbors, an $N_{ne} \times H$ binary matrix A was constructed, where the element A_{ij}^h is equal to 1 if protein p_i has an interacting neighbor p_j located in the adjacent SCL l_h and 0 otherwise. μ_h is an unknown parameter for the adjacent SCL l_h. The parameters $\alpha, \eta, \beta^{11}, \beta^{10}$, and μ are estimated during optimization.

4.1.2 Gibbs sampler and likelihood estimation

Energy functions are the negative logarithm of the posterior probability distribution of the SCL labeling. Maximizing the posterior probability equals to minimizing the energy function, which is defined as $\mathbf{x} = argmin_{x \in L} E(\mathbf{x})$. In this study, the approximation method maximum pseudo-likelihood estimation (MPLE) was used to solve the maximization problem. The general idea of this approach is to learn model parameters by maximizing the pseudo-

likelihood, which replaces the likelihood with a tractable product of conditional probabilities [145, 61]. Since the SCL datasets are usually highly imbalanced, the posterior $Pr_\theta(x_i|\mathbf{x}_{-i})$ will tend to be overwhelmed by the majority classes (in this case negative examples in individual binary classifier). In order to deal with this problem, an imbalance coefficient is used to re-balance the influence on the joint likelihood by enhancing the minority classes [132]. Thus the re-balanced pseudo-likelihood function (PLF) can be written as

$$PLF(\mathbf{x}) = \prod_{i=1}^{N} (Pr(x_i|\mathbf{x}_{-i}))^{c_i^m} \tag{4.8}$$

where c_i is the imbalance coefficient

$$c_i^m = \begin{cases} \frac{n^-}{n^+} & x_i = 1 \\ 1 & x_i = 0 \end{cases} \tag{4.9}$$

where n^+ and n^- denote the numbers of positive samples and negative samples for the SCL l_m, respectively.

Assuming that the parameter set θ is given, for a given protein conditional on the SCL labels of all of the other proteins $Pr_\theta(x_i|\mathbf{x}_{-i}) \approx Pr_\theta(x_i)$. Therefore, we can use MPLE with $Pr_\theta(x_i|\mathbf{x}_{-i})$ to generate samples to update the SCL labels of protein p_i as follows:

$$Pr_\theta(x_i = 1|\mathbf{x}_{-i}) = \frac{P_\theta(x_i = 1, \mathbf{x}_{-i})}{P_\theta(x_i = 1, \mathbf{x}_{-i}) + P_\theta(x_i = 0, \mathbf{x}_{-i})} \tag{4.10}$$

with

$$\begin{gathered} Pr_\theta(x_i = 1, \mathbf{x}_{-i}) = \frac{1}{Z(\theta)} \exp(-E(x_i = 1, \mathbf{x}_{-i})) \\ \text{and } Pr_\theta(x_i = 0, \mathbf{x}_{-i}) = \frac{1}{Z(\theta)} \exp(-E(x_i = 0, \mathbf{x}_{-i})) \end{gathered}$$

$$Pr_\theta(x_i = 1|\mathbf{x}_{-i}) = \frac{1}{1 + exp(E(x_i = 1, \mathbf{x}_{-i}) - E(x_i = 0, \mathbf{x}_{-i})}) \tag{4.11}$$

Given θ, the conditional probability of x_i has the SCL given its neighbors

$$Pr_\theta(x_i = 1|\mathbf{x}_{-i}) = \frac{1}{1 + e^{v_i}} \tag{4.12}$$

where

$$\begin{aligned} v_i = \alpha + (\beta^{11} - \beta^{10}) K_i^1 + \beta^{10} K_i^0 + \\ \sum_{h=1}^{H} \mu_h K_i^h + \phi_i^F(x_i, F_i) \end{aligned} \tag{4.13}$$

Algorithm 2: Gibbs sampling

Data: PPIN partially annotated with SCL; Temperature: T; Cooling rate: R. Maximum iteration: I. Cooling iteration: I_c.

Result: Fully annotated PPIN with predicted SCL which associated with probability values.

```
Initialize the parameter set θ_c using linear logistic regression based on known proteins.
Initialize the x_i value of unlabeled proteins.
Calculate PLF(x, θ_c)
Sample θ_p ← θ_c
while I_c <= I do
    I_c = I_c + 1
    Sample θ_p ← θ_c, Equation (4.16);
    Calculate PLF(x, θ_p), Equation (4.8);
    Calculate acceptance Probability r = P(PLF_c, PLF_p, T), Equation (4.17);
    if r > r_unif then
        θ_c = θ_p;
        Update the value of x_i based on Equation (4.12);
    end
    T = T · (1 − R)
end
```

K_i^1 is the weighted number of neighbors of protein p_i which are assigned to the SCL, and K_i^0 is the weighted number for neighbors of p_i which are not. Likewise, $K_i^h{}_{,h\in H}$ is the number of neighbors of p_i which are assigned to the adjacent SCL l_h factorized by S_h and ω_{ij}. x_{-i} is the set of proteins without the i-th protein. P_i^F is the probabilities that protein p_i locates in the SCL l_m depends on this feature. P_i^A is the probability that protein p_i locates in the SCL l_m depends on if the protein interact or not with proteins in its adjacent SCLs.

And

$$Pr_\theta(x_i = 0|\mathbf{x}_{-i}) = 1 - Pr_\theta(x_i = 1|\mathbf{x}_{-i}) \tag{4.14}$$

Repeating this procedure many times generates samples for the SCL of all of the unannotated proteins. Considering the computational complexity, simulated annealing searching algorithm was applied to find the local optimal solution, see Algorithm 2.

The PLF is the product of the conditional probabilities across all proteins using Equation (4.12) and Equation (4.14).

$$PLF_\theta(\mathbf{x}) = \prod_{i=1}^{N} Pr_\theta(x_i|\mathbf{x}_{-i}) \tag{4.15}$$

4.1.3 Parameter learning

The estimation of the parameters is realized by maximizing the PLF in Equation (4.8).

Parameter initialization

$Pr_\theta(x_i = 1|\mathbf{x}_{-i})$ is a logistic function on a linear function (see Equation (4.12) and Equation (4.13)). With the sub-network of all of the annotated proteins, the parameter set θ can be initiated using linear logistic regression based on the SCLs of these annotated proteins.

Parameter update

Instead of sampling each parameter of θ separately, for each iteration a parameter set θ_p was estimated using MCMC DEMC algorithm [146, 61].

$$\theta_p = \theta_c + \gamma(Z_1 - Z_2) + e \tag{4.16}$$

where θ_c denotes the current state of the parameter vector, γ follows the uniform distribution $U(\gamma^*/2, \gamma^*)$ is the scaling parameter and $\gamma^* = \frac{2.38}{\sqrt{2d}}$ is the optimal step size, where d is the parameter dimension. Z_1, Z_2 are uniformly selected from past samples of the Markov Chain as stored in a matrix Z and $e \sim$ Multivariate Normal Distribution $MVN(0, 10^{-4})$. θ_p is accepted using a metropolis step with simulated annealing:

$$r = \exp \frac{PLF(x^t|\theta_p) - PLF(x^t|\theta_c)}{T} > r_{unif} \tag{4.17}$$

where $r_{unif} \sim unif(0,1)$, T is the temperature initialized with 150 and the cooling rate is 0.008 per iteration.

4.1.4 Collective MRFs

In the MRFs, each variable x_i in vector $\mathbf{x} = \{x_1, \ldots, x_N\}$, represents whether a protein being located to l_m or not. For protein p_i, it is possible that its neighbors located in adjacent SCLs are also unknown. To respect the property of MRF, the labels of unknown proteins were initialized by the labeling results from the MRFs model without considering the adjacent SCLs. The results from the previous MRFs $MRF - l_m^t$ are collected and used in the next MRFs, such as $MRF - l_{m+1}^t, \ldots, MRF - l_M^t$. This process is repeated iteratively until the convergence of the pseudo-likelihood Equation (4.8). These MRFs are therefore named as BCMRFs (see Figure 4.2 and Line 13). BCMRFs approach is similar to the α-expansion algorithm [147], which is a popular for solving the multi-label MRFs in imaging processing.

Algorithm 3: Collective MRFs

Input: M partial labeled network for the M SCL terms
Output: M fully assigned network for the M SCL terms

```
for each SCL terms do
    Initialize the x_i values of unknown proteins using MRF model without adjacent
     SCLs potential φ^A (Equation (4.7)).
end
while PLF(x) not converge do
    for each SCL terms do
        Optimize the PLF^t(x).
        Calculate acceptance probability r comparing with the PLF^{t-1}(x).
        if r > r*_unif then
            Update the labeling of x according to PLF^t(x).
        end
    end
end
```

∗: r_{unif} is a random variable follows uniform distribution.

4.1.5 Computational complexity

Given the number of SCL m, number of the protein in the PPIN n, the estimation of the running time of BCMRFs $f(n)$ can be written as follows:

$$f(n) = m \cdot I_1 \cdot g(n) \tag{4.18}$$

with maximum iteration limit I_1 in which the optimization can converge, and the running time of Algorithm 2 $g(n)$:

$$g(n) = \begin{cases} (c_1 + c_2 \cdot I_2) \cdot n + (c_3 + I_2) \cdot n^2 & \text{best case} \\ (c_1 + c_2 \cdot I_2) \cdot n + (m + (m + c_3) \cdot I_2) \cdot n^2 & \text{worst case} \end{cases} \tag{4.19}$$

with the costs c_1, c_2, c_3 and maximum iteration limit I_2. The best case is that the number of proteins which SCL are unknown $n_{unknown}$is as small as 1, number of proteins which SCL are known is as close to the total number of proteins n. The number of adjacent SCCs n_{adj} and number of features n_{feat} are zero, whereas the worst case is $n_{unknown} \approx n$, $n_{known} = 1$, $n_{adj} = m$, and $n_{feat} = m$. Since c_1, c_2, c_3, I_1, I_2 and m are constants, the estimated computational complexity is therefore $\mathcal{O}(n^2)$.

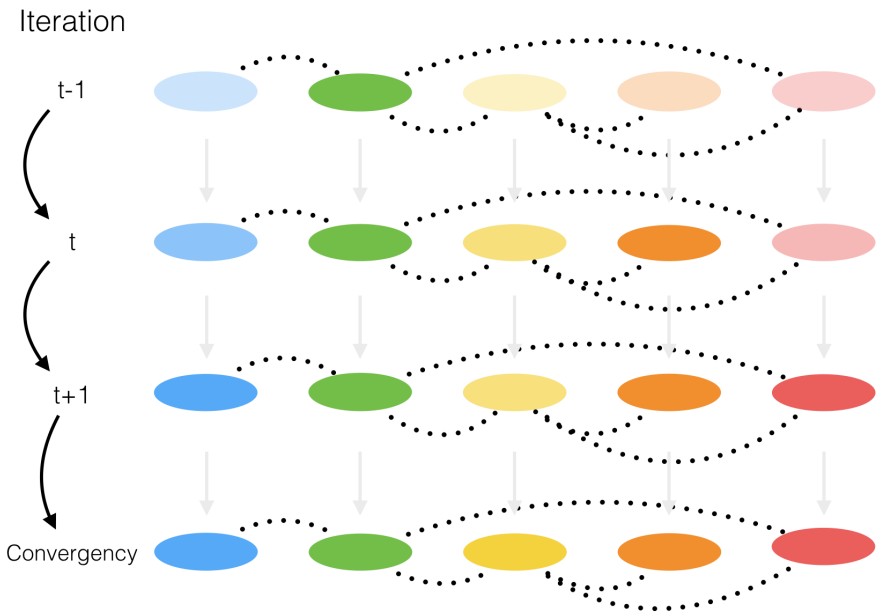


Fig. 4.2 Overview of the collective MRFs. Each colored eclipse represents a MRF for one SCL term. The different shades of color represent the joint likelihood value. The deeper the color, the higher the likelihood calculated from this MRF. The dotted line between eclipses represents the spatial relationship of SCLs.

4.1.6 Implementation

The BCMRFs algorithm is implemented in R language. The source code can be found at `https://github.com/zhu0619/BCMRFs`. The program is mainly divided into two parts. 1. Preprocessing part contains scripts which either generate the protein SCL dataset, acid amino sequences, and the PPI dataset from a list of proteins or import customized datasets and process the datasets to match the format of BCMRFs program. 2. Main programs process the BCMRFs analysis on the given partially annotated PPIN and return the predicted SCL results of the protein which were unknown. The results are associated with posteriori probabilities. An overview of the implementation and program parameters is given in Figure 4.3.

4.2 Experimental setup

4.2.1 Dataset

A recently published high-quality human protein SCL benchmark set from the subcellular localization database Compartments [148] was used to evaluate the performance of BCMRFs method. In total nine SCLs including Cytosol, Endoplasmic Reticulum, Lysosome, Extracel-

lular space, Golgi apparatus, Mitochondrion, Nucleus, Peroxisome and Plasma membrane are used for evaluation. The dataset was created from UniProtKB and HPA.

The corresponding protein sequences for generating the features from YLoc were retrieved from UniProtKB (version 2016.08). The PPI data were retrieved from the interactom browser - Mentha [49] (version 2016.09). It limits itself to direct physical PPIs curated by members of the International Molecular Exchange consortium (IMEx) [45]. Each PPI is associated with a reliability score which takes the evidences such as experimental method, size of experiments and relevant literature into account [49]. In Figure 4.5, there is a dramatic reduction of PPIN size with a cutoff of reliability score 0.25. It indicates that most of the low quality PPIs in the network can be removed by using this cutoff value. The remaining PPIs are weighted by the reliability scores for the MFRs. In the filtered connected PPIN, 5496 proteins are SCL-known while 1299 protein have no SCL annotation available. Figure 4.4 further shows the distribution of the SCLs of our human proteins data set. As can be seen, of the 5496 proteins, 4367 are single-SCL located proteins, 1129 have from two to seven SCL annotations. As shown in the pie chart, the almost 50% of of single-SCL proteins locate in the nucleus which is consistent with the distribution of the overall proteins. For the proteins locate in two or more SCLs, nucleus shows less and less portion in the distribution. Therefore, the single-SCL protein plays more significant roles in shaping the overall distribution of the data set. Nevertheless, the multi-SCLs proteins which takes big percentage of the population can not be ignored.

4.2.2 Evaluation

To evaluate the prediction performance of the proposed method, a six fold cross validation was performed. For 1000 out of 5496 proteins, their SCL labels were masked, and treated as unknown protein. Hence, 2299 proteins in the network are unlabeled. And the predicted label of these masked protein are used for performance evaluation. The dataset stratification was done by using R package "utiml" [149].

The traditional performance measures are difficult to apply for multiple SCL prediction. To better reflect the multi-label capabilities of classifiers, the popular multi-label measures were used including Precision (PRC), Recall (RCL), F-measure (F1 score), Average Precision (AP) and Hamming Loss (HL) [149]. Except HL, for all the rest of performance measures, the higher the measures, the better the prediction performance. To keep the consistency, the 1-HL was shown instead of HL for the evaluation.

$$Accuracy = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup |Z_i|} \tag{4.20}$$

$$Precision = \frac{1}{D}\sum_{i=1}^{D}\frac{|Y_i \bigcap Z_i|}{|Z_i|} \tag{4.21}$$

$$Recall = \frac{1}{D}\sum_{i=1}^{D}\frac{|Y_i \bigcap Z_i|}{|Y_i|} \tag{4.22}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.23}$$

$$Hamming\ loss = \frac{1}{|\mathbf{D}|}\sum_{i=1}^{|\mathbf{D}|}\frac{Y_i \triangle Z_i}{\mathbf{L}} \tag{4.24}$$

4.2.3 Comparison partners

In the experiments, in order to investigate the effects of different potentials described in section Section 4.1, four versions of MRFs which include different combinations of potentials, such as MRFs with PPI only (M1) ,with PPI and SCL spatial adjacency (M2), with PPI and protein features (M3), and the MRFs with all three defined as Equation (4.2) (M4) were built and compared.

Moreover, the MRFs were also compared with state-of-art SCL prediction methods, including:

- DC-*k*NN proposed by [108] provides the best SCLs predicting result for human proteins based on PPIs. In their study, they reported the SCLs for 4366 human proteins with no SCL previously known at the time in 2008 predicted by their method. From then to 2016, 1704 of these proteins has been reported in various SCLs. The SCL annotations were collected following the same criteria as their benchmark [108].

- Hum-mPLoc 3.0 is a most recent protein feature-based SCL predictor for human proteins [97]. The predicted SCLs of 5390 human proteins from their database are used for the comparison.

4.3 Results

4.3.1 Likelihood and prediction performance

In the proposed method, the protein SCLs were predicted by minimizing the energy function. In other words, the higher the calculated conditional probability of a protein given its interacting neighbors for a certain SCL l_m, $1 \leq m \leq M$, the more confident that this protein locates in l_m, which infers that the overall prediction performance achieves for l_m should be positively correlated with the data likelihood. Figure 4.6 shows that the lower the energy (the negative logarithm of the likelihood) is, the higher the F1 score is which confirms the concept.

4.3.2 Effects of different potentials

To investigate the effect of the potential described for the prediction, the performances including of the four versions of MRFs M1, M2, M3, M4 were compared.

Single-SCL prediction

At first, the performance of the four models for each SCL class were compare individually. **M2 VS M1** : Figure 4.9a shows that the spatial SCL adjacency relation of interacting proteins can improve the prediction for the majority of the SCL classes, except Lysosome and Peroxisome, which even the decrease in prediction performance. Firstly, these two SCL classes are highly imbalanced with few positive labels (see Figure 4.8). Moreover, the prediction on the SCL Cytosol is quite poor. Therefore, the MRFs of Lysosome and Peroxisome can not gain the correct information from their only spatial adjacent SCL Cytosol to increase their prediction performance. In order to put the spatial adjacency to good use, it is necessary to firstly improve the overall performance. Therefore, it is necessary to integrate the potential based on protein features into MRF model (M3). With regard of Cytosol, it is an intracellular fluid which comprises most of the cellular organelles, and involved in many biological processes. The low performance could be due to its complexity. The features can not improve the prediction performance. Finally, adding the SCL adjacency potential above on M3, the prediction performances were improved on most of the SCL classes.

Multi-SCLs prediction

As can be seen from Figure 4.9b, M2 outperforms M1 which means additional spatial adjacency can improve the performance comparing with the simple SCL inference based on

PPI only. However, the improvement is limited due to that M2 cannot efficiently gain correct knowledge from the adjacent SCLs which are poorly predicted. As expected, M3 significantly improve of prediction performance by adding the features of proteins on the model of M1. M4 can achieve the best performance of all. Comparing with M3 in particular, together with the observations of single-SCL predictions, it can be concluded that the improvement is owing to that the model can efficiently gain the correct knowledge from the adjacent SCLs. However, in order to show a larger improvement of performance of the multi-SCLs prediction by adding the spatial adjacency on the proteins features in the model (M4 against M3), an additioanl tuning of parameters would be necessary.

4.3.3 A collective process improves the performance

To demonstrate how the collective MRFs can help to improve the performance of the SCL prediction, the changes of performances of M4 during the 21 iterations is shown in Figure 4.7. Overall, the F1 scores gradually increase from initialization (iteration 1), single MRFs (iteration 2) and collective MRFs (from the 3^{rd} iteration). The performances stay stable as the pseudo likelihood value of BCMRFs converge.

4.3.4 Transductive learning from imbalanced MLDs

The human protein dataset is highly imbalanced since some of the labels are very frequent whereas most others are rarely used. The imbalance level of a MLD can be effectively measured by the imbalance ratio (*IRLbl*) [64]. Figure 4.8 shows that the SCLs such as Lysosome and Peroxisome are highly imbalanced compared to the other SCLs, with IRLbl of 22.4 and 44.18 respectively.

Facing the imbalance problem, the popular solution is data resampling including under-sampling and over-sampling [64]. However, in the case of SCL prediction, the re-sampling techniques cannot be applied due to the proposed method being highly sensitive to the topology of PPIN. The inference of SCL in this approach depends on the number of physical interactions. Under-sampling and over-sampling are based on the deletion of true interactions or repetition of existing interactions which can largely change the topology of the network and thus mislead the MRFs. Therefore, in this study the imbalanced MLD problem was handled by introducing imbalance coefficient (see Equation (4.9)). The prediction performances of the BCMRFs with and without the imbalance coefficient was compared. The results in Table 4.1 shows that the MRFs with the imbalance coefficient can improve the performance.

Furthermore, a comparison of the prediction results of the BCMRFs built on the complete PPIN including the unknown proteins against the BCMRFs built only on the sub-network

Table 4.1 F1 scores with/without imbalance correction.

Model	M1	M2	M3	M4
With imbalance coefficient	0.637	0.641	0.71	**0.732**
Without imbalance coefficient	0.616	0.632	0.701	0.722

Table 4.2 F1 scores for transductive VS conventional.

Model	M1	M2	M3	M4
Transductive learning	0.648	0.652	0.743	**0.759**
Conventional learning	0.602	0.647	0.684	0.692

of the annotated proteins was performed. Table 4.2 shows that the MRFs of transductive learning outperforms the MRFs of the conventional learning.

4.3.5 Comparison with existing methods

To further demonstrate the performance of the method, the BCMRFs was compared with the only available PPI-based approach for predicting human protein SCLs, DC-*k*NN [108] and the protein feature-based method Hum-mPLoc 3.0 [97]. DC-*k*NN is a physical PPI-based prediction method using a *k*-nearest neighbors classification with binary reference approach. Due to the unavailability of the program and of its prediction results, the dataset which were used to compare BCMRFs methods only contains 1704 human proteins (see Section 4.2.3). For these 1704 human proteins, an evaluation of the prediction results of DC-*k*NN and the results of BCMRFs method was carried out. Table 4.3 shows that BCMRFs method significantly outperforms DC-*k*NN overall.

Hum-mPLoc 3.0 [97] is the state-of-the-art feature-based SCL predictor specifically for human proteins. It predicts SCLs based on the amino acid sequence of proteins through modeling the hidden correlations of gene ontology and functional domain features. The comparison of multi-SCL prediction results from Table 4.4 demonstrate that BCMRFs method achieves better performance.

4.4 Summary

Protein SCL prediction is an imbalanced multi-label classification problem. This chapter described a bayesian collective MRFs algorithm to predict multi-SCLs of human proteins. This is done by building the weighted MRFs based on the PPIN and then performing SCL

Table 4.3 Comparison with the method of DC-*k*NN - Multi-SCL prediction.

Method	Precision	Recall	F1 score	Average precision	Hamming loss
DC-*k*NN	0.502	0.472	0.474	0.672	0.119
BCMRFs	**0.674**	**0.621**	**0.633**	**0.899**	**0.073**

Table 4.4 Comparison with the method of Hum-mPLoc 3.0 - Multi-SCL prediction.

Method	Precision	Recall	F1 score	Average precision	Hamming loss
Hum-mPLoc 3.0	0.68	0.688	0.660	0.735	0.090
BCMRFs	**0.702**	0.67	**0.673**	**0.862**	**0.078**

label propagation to predict the SCLs of unknown proteins. The comprehensive experiments were performed to evaluate the performance on human protein SCL datasets. The transductive learning from the re-balanced MLD proved to be more efficient to correctly assign SCLs. Owing to the collective MRFs which connect the binary MRFs by their spatial adjacency among SCLs, BCMRFs can achieve a higher performance for predicting the multi-SCLs comparing with the state-of-the-art methods of DC-*k*NN and Hum-mPLoc 3.0.

Interestingly, neither the present approach nor the previous state-of-the-art method for SCL prediction perform as effectively for human as for other organisms (such as bacteria: precision > 0.95 and recall > 0.93 for single-SCL prediction) [150]. One explanation could be that the cell structures of the bacteria (5 and 6 SCC in total) are much simpler than mammalian cells. The the activities of human cells, such as the interactions among proteins and with other molecules, the translocalization of proteins, the functions of proteins, and the biological environment of the cell are also more complicated. Therefore, there may still be room for improvement of the SCL prediction of human proteins.

All PPI data used in this task are static data reported from different studies and techniques with a huge diversity. During different biological processes, one protein can play different roles and functions, for instance by interacting with different target proteins. However, the available PPI datasets do not differentiate them according to the biological contexts. Since a single protein cannot physically interact with tens or hundreds of partners at the same time, this presents a future challenge: How to determine which interactions occur simultaneously and which are mutually exclusive? And how to explore this knowledge to make tissue-specific SCL predictions? The answers are presented in the next chapter, see Chapter 5.

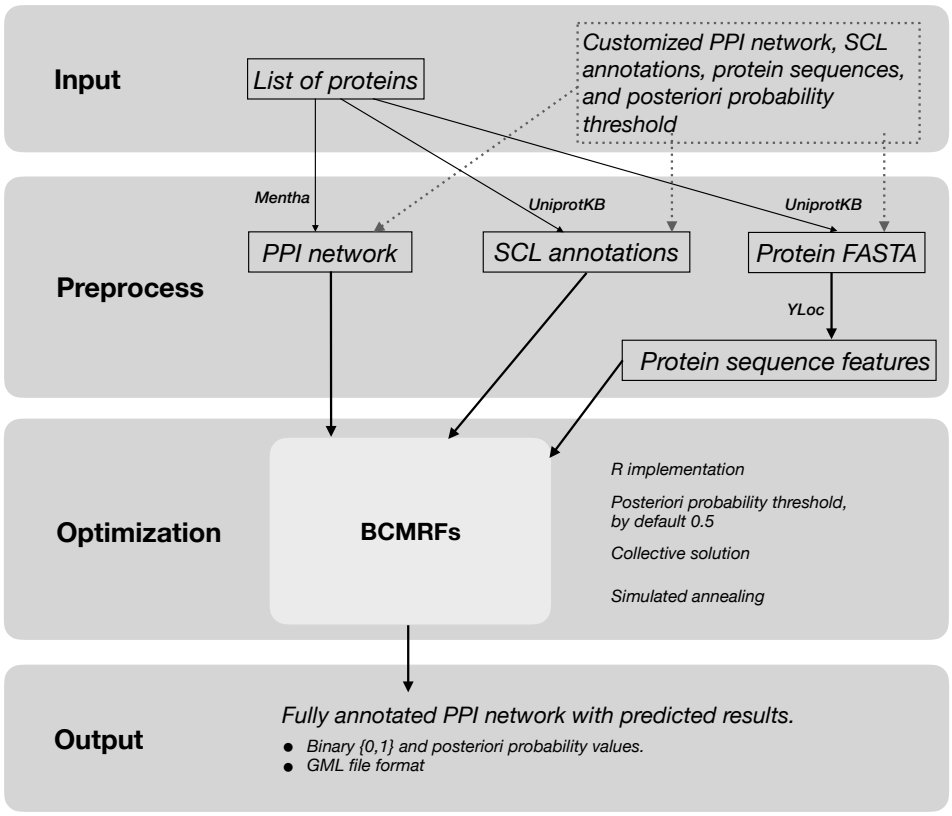


Fig. 4.3 The overview of implementation of BCMRFs method. The proposed method is implemented in R and divided into the preprocessing and the optimization part. The R scripts for the preprocessing compute Furthermore, the user can also customize the PPIN, the SCL annotation, the protein sequences according to their research of interest. All the dataset are then preprocessed to match further computing. User can also improve precision or recall by setting the cutoff value for the posteriori probability value. As output a Geography Markup Language (GML) formated file is provided for further visualization and modeling.

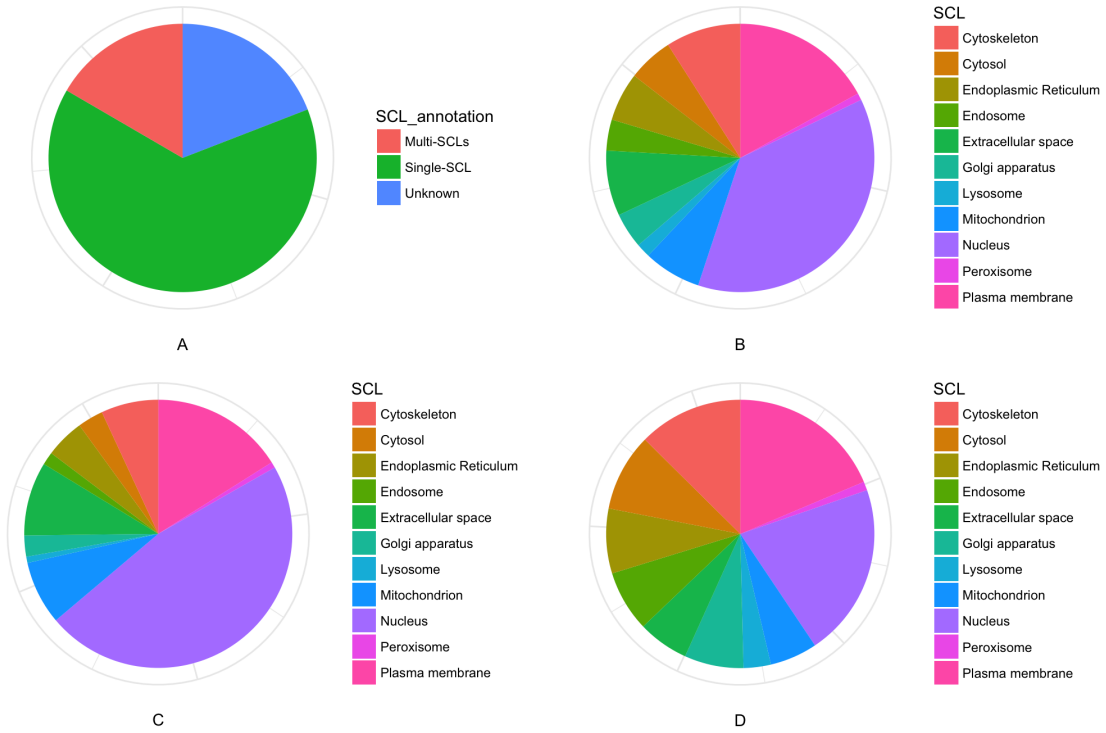


Fig. 4.4 Summarization of descriptive data from the human protein dataset. A. SCL annotation of proteins; B. Overall distribution of protein in SCL classes; C. Distribution of single-SCL protein; D. Distribution of multi-SCLs proteins.

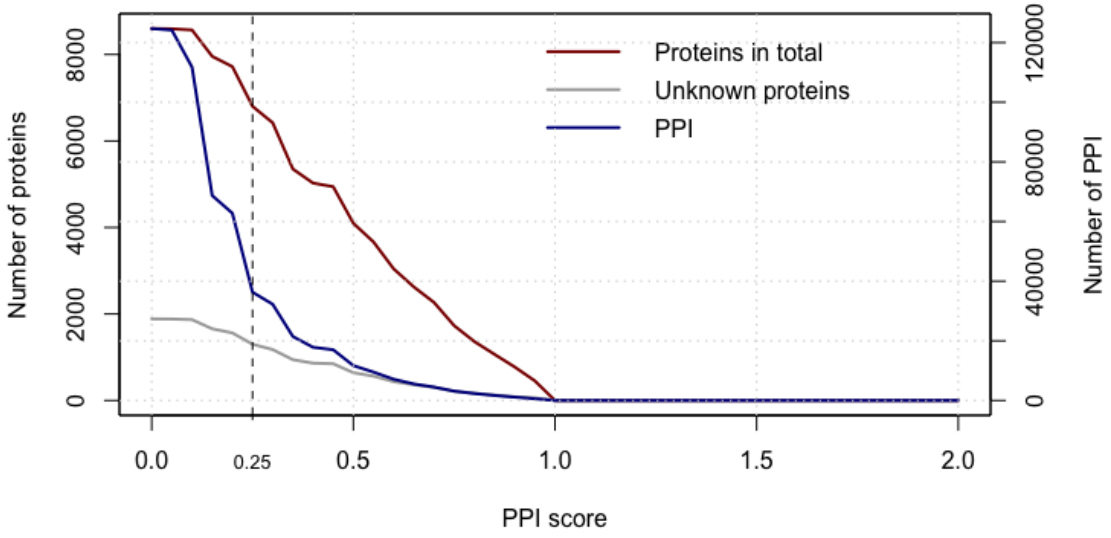


Fig. 4.5 Protein-protein interactions of test dataset controlled by the confidential scores.

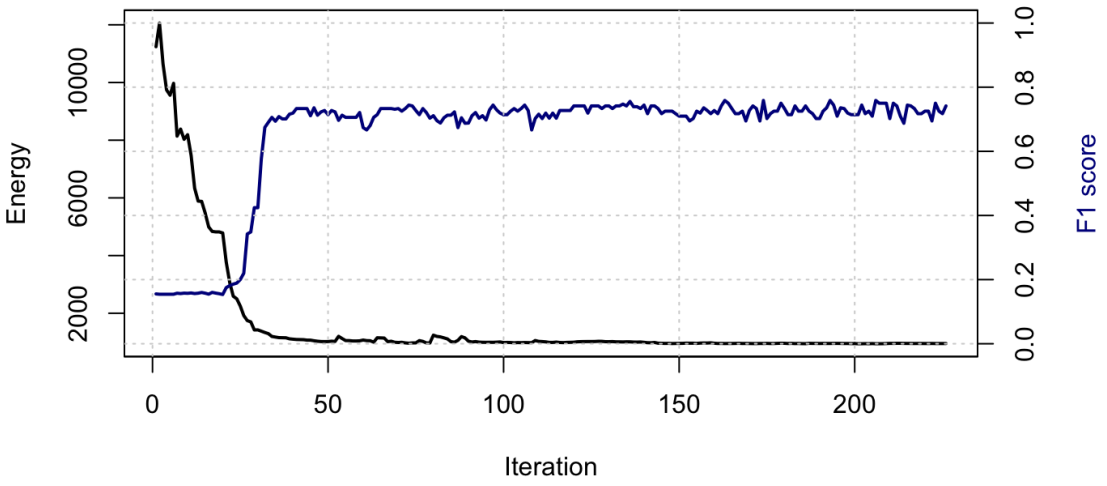


Fig. 4.6 Relationship between the likelihood and prediction performance.

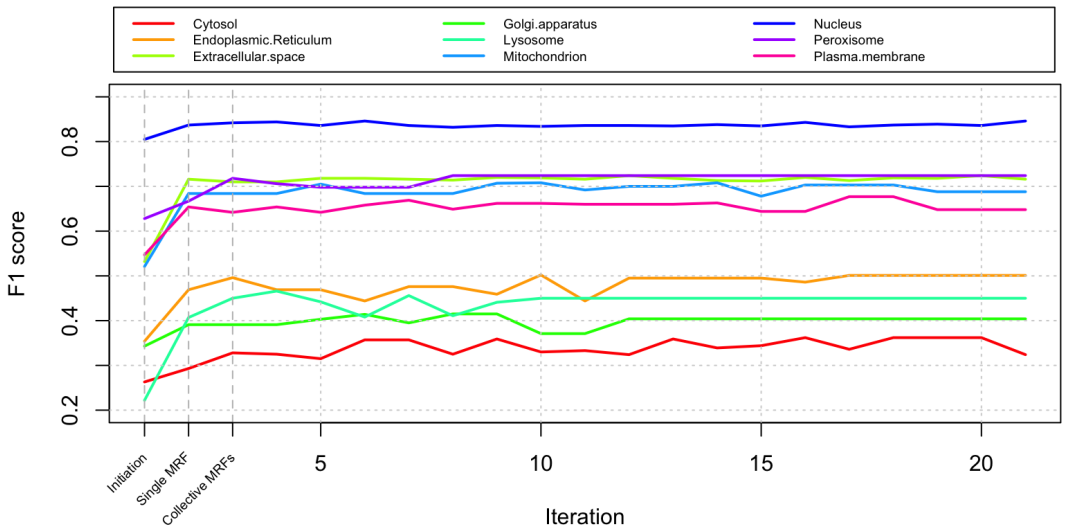


Fig. 4.7 Performances of BCMRFs during iterations.

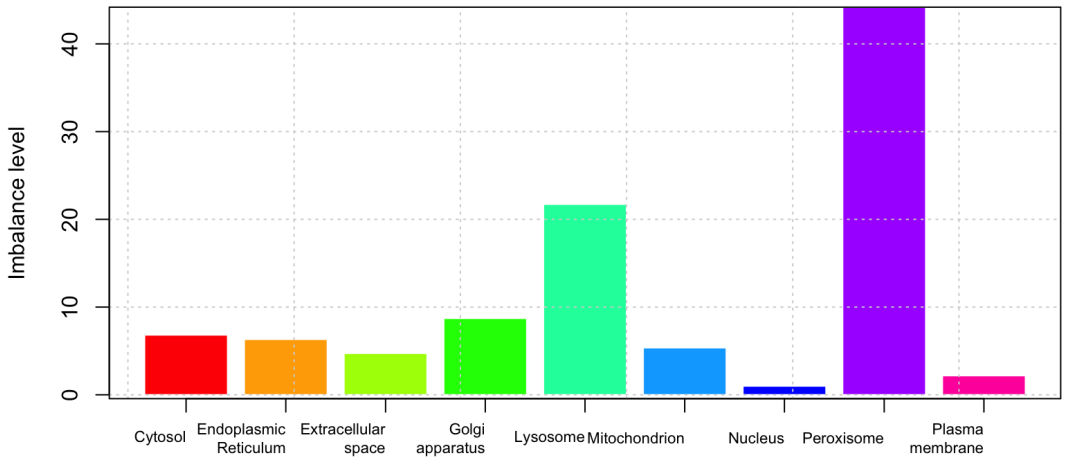
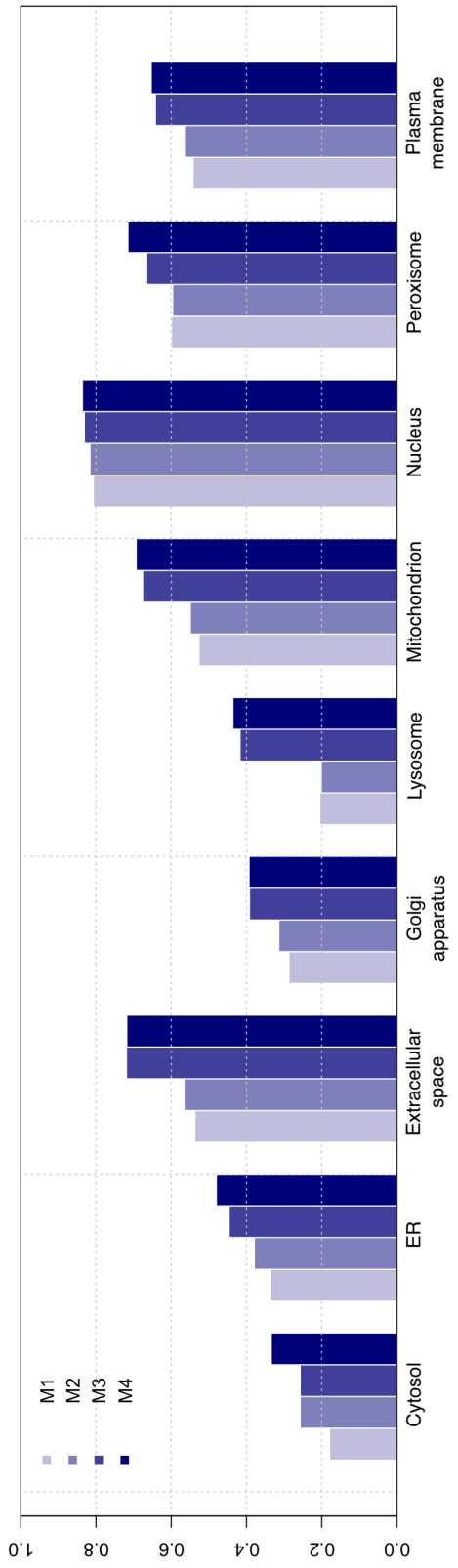
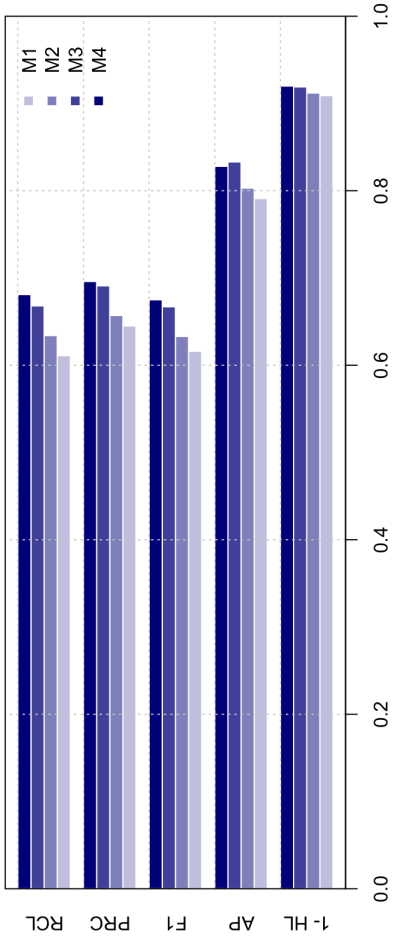


Fig. 4.8 Imbalance level of each SCL class.



(a) Performance of the single-label classification.



(b) Performances of the multi-label classification.

Fig. 4.9 Prediction performances of four models.

Chapter 5

Tissue-specific SCL prediction

Proteins interact with each other or with other types of molecules to carry out the specific functions and the dynamic activities, such as form cellular machines and to pass signals within cells and across tissues. To interact, proteins (or any other molecules) must necessarily share a common SCC, or an interface between physically adjacent SCCs, transiently or conditionally. Hence, the SCL of a protein can therefore be inferred from the SCL of its interacting partners. Previous chapter described the algorithm BCMRFs which uses physcial PPINs for protein SCL prediction.

However, the character of the generic PPI data shows that one protein can be observed having tens to hundreds interacting partners which apparently exceeds the physical limits of a protein's contact surface. The reason is that these data are collected and accumulated from different studies and a diverse ranges of biological contexts.

Thus, some of the generic PPIs can be considered as 'false positive' interactions for one biological context to another. For instance, an interaction which can only occur in brain tissue would not appear in skin tissue. Hence, an SCL of a protein in brain tissue should not be inferred from the SCL can be only observed in skin tissue of that protein using the skin tissue specific PPI .

While knowledge of context-specific PPIs is limited, there is a rapid accumulation of context-specific molecular expression profiles. Many studies revert to identifying PPIs that are feasible in these contexts with the assumption that a PPI is possible within a specific context if the corresponding proteins are expressed in that context, such as disease-specific PPIs [151] and tissue-specific PPIs [10, 137]. The tissue-specific PPIs of human protein have been intensively studied and well explored [9]. Among all of the research on context-specific PPI, the data of tissue-specific PPI have been well studied last decade.

Similar to PPI data, the lack of context-specificity also exists in protein SCL data that were measured in different tissues and cell lines. The HPA has reported that more than

50% of the analyzed proteins in their database were identified to localize to more than one compartment at the same time. These MLP may have context-specific functions increasing the functionality of the proteome. Among these MLPs, 3546 proteins showed cell line dependent properties [28]. These cell lines were derived from corresponding tissue types and often used as model systems in human biology and diseases.

Furthermore, some proteins indeed show their tissue specificity at the subcellular level. First of all, some particular SCCs exist only in specific tissues. For instance, the sarcolemma is a unique subcellular compartment in muscle tissue. Moreover, the spatial distribution of the SCLs of the protein in a cell could be different from one tissue to another, which depends on the functions of the protein in the specific tissue. For example, glutamine synthetase (GS) is mitochondrial in liver cells and cytoplasmic in brain cells [15]. In the human tissue adrenal gland, pituitary gland and pancreas, the absence of adracalin (ALADIN) in nuclear membrane causes human triple A syndrome [16]. Therefore, understanding the specific SCLs of human proteins in different tissues and organs of the human body would significantly increase our knowledge of human biology and disease.

Despite the growing understanding of the tissue-specific proteome, to the best of our knowledge, there is not yet a computational method for predicting tissue-specific SCLs. For this purpose, a new approach using a probabilistic graphical model integrating tissue-specific functional associations and physical PPINs to predict tissue-specific SCL are described in this chapter.

To reveal a landscape of dynamic changes in SCLs across tissues, the previously introduced BCMRFs model can be applied to predict tissue-specific SCLs by integrating tissue-specific physical PPIN. The rest of this chapter is organized as follows: In Section 5.1 we remind our BCMRFs model, the corresponding learning procedure, the experimental protocol and data sources. Section 5.2 shows the statistics and the analyses on both tissue-specific physical PPI and tissue-specific SCLs of human proteins, and the experimental results. At last, we summarized our work and the discussion of directions for future work.

5.1 Methods

At first, the general ideas of the method for tissue-specific SCL prediction are introduced, and followed with the detailed description of the MRFs model and the data sources which were used in this study.

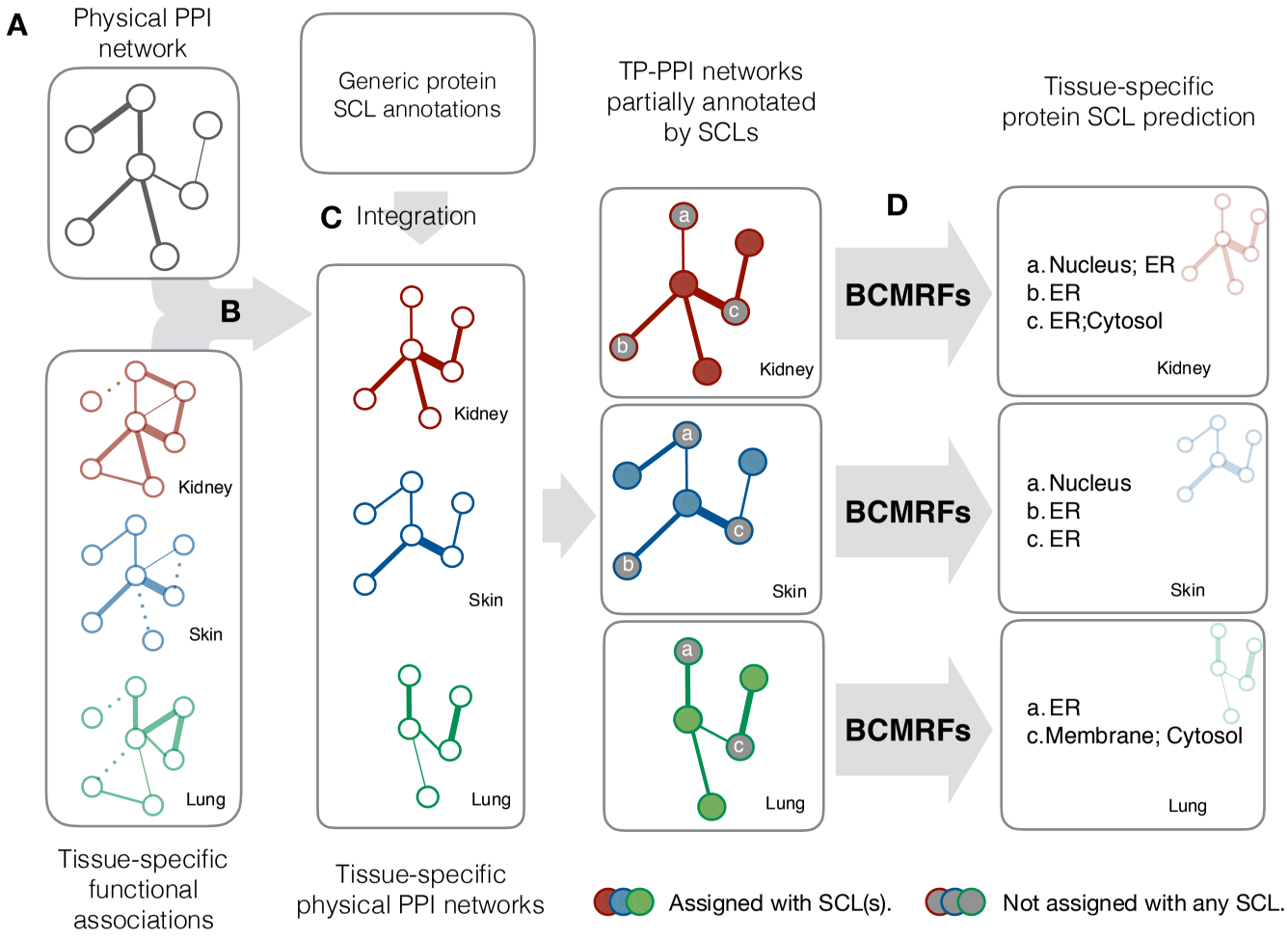


Fig. 5.1 The workflow of the tissue-specific SCL prediction based on PPINs. A. Integration of generic physical PPIN and tissue-specific functional PPIN; B. Construction of the tissue-specific protein-protein interaction networks; C. Integration of the SCL annotations with each tissue-specific PPINs; D. tissue-specific SCL prediction using BCMRFs based-on tissue-specific PPINs.

5.1.1 BCMRFs for predicting tissue-specific SCLs

As illustrated in Figure 5.1, to perform tissue-specific SCL prediction, our first step is to construct a tissue-specific PPIN. Both the generic physical PPINs and tissue-specific functional associations were integrated (Figure 5.1 A). Although the tissue specificity is demonstrated by the functional association of pairwise proteins, only the physical PPIs provide the direct evidence of SCL dependency. Thus, the generic physical PPINs were filtered using the tissue-specific functional associations to construct tissue-specific PPINs for each tissue (Figure 5.1 B). Afterwards, each TP-PPI network is annotated by the generic protein SCL annotations (Figure 5.1 C). Due to the tissue-specific expressed proteins are different from one tissue to the other, the topology of the tissue-specific PPINs are also different. For each tissue-specific PPIN, a corresponding multi-label MRFs was built to predict tissue-specific multiple SCLs of unknown proteins by inferring from the interacting proteins of which the SCL are already known. The tissue-specificity of protein SCLs depends on the topology and the edge weights of the tissue-specific PPINs.

In a protein SCL dataset, a protein is generally associated with a set of SCL labels which makes the protein SCL prediction a multi-label classification problem. Previously, Chapter 4 introduced BCMRFs to predict the multiple SCLs of proteins based on a physical PPIN. The general idea is to solve the multi-label MRF by reducing it to multiple binary MRFs using the binary reference approach. The final optimal solution for the mutli-label MRFs can be obtained by merging the solutions from binary MRFs.

The only modification compared to the MRF model for generic SCL prediction is the edge weight parameter. In addition to the reliability score of the physical interaction, the tissue-specific functional associations was integrated that each pair of PPI is associated with the probability of the tissue-specificity. The final edge weight is the product of tissue-specific probability and the physical interaction reliability score. In the following, the definition of MRF model is reminded.

The posterior distribution $Pr(\mathbf{x})$ over the SCL labelings of the muti-label MRF is a *Gibbs* distribution which is defined as below:

$$Pr(\mathbf{x}) = \frac{1}{Z}\exp(-E(\mathbf{x})) \tag{5.1}$$

where Z is a normalizing constant known as the partition function. $E(\mathbf{x})$ is the energy function of the MRFs. A series of changes are made to the current solution to decrease its energy. A set of label changes is called a move. In one iteration of the algorithm, it makes moves with respect to each SCL label 'm' ($\in \mathscr{L}$). It finds the optimal changes to be made by minimizing a binary energy function. The binary energy function corresponding to a particular 'm' move

will be denoted by $E^m(\mathbf{x}^m)$. It is defined

$$\begin{aligned} -E^m(\mathbf{x}^m) = &\sum_{i \in v} \phi_i^m(x_i^m) + \sum_{i \in v} \phi_i^{Fm}(x_i^m, F_i^m) \\ &+ \sum_{i,j \in \varepsilon} \omega_{i,j} \phi_{ij}^{Pm}(x_i^m, x_j^m) \\ &+ \sum_{i,j \in \varepsilon} \omega_{i,j} \phi_{ij}^{A}(x_i^m, x_j^m, A_{ij}^m) \end{aligned} \tag{5.2}$$

where $\omega_{i,j}$ is a constant parameter. It is the product of the confidential score of the physical PPI and the probability of tissue-specificity of the functional associations between P_i and P_j.

Moreover, the estimated computational complexity is also $\mathcal{O}(n^2)$, with running time $e(n) = t \cdot f(n)$ while t is number for tissues, $f(n)$ see Section 4.1.5. The detailed method and formulas are described in Chapter 4 Section 4.1.

5.1.2 Implementation

The implementation tissue-specific BCMRFs method was based on the implementation of BCMRFs which were described in previous chapter. Additional R scripts were implemented to retrieve tissue-specific functional associations via API and generate tissue-specific physical PPIN. An overview of the modified implementation is given in Figure 5.2. The source code and an example can be found at `https://github.com/zhu0619/BCMRFs`.

5.1.3 Data resources

SCL of human protein datasets

Generic SCL dataset In this study, the generic SCL annotations of human proteins are retrieved from the SCL database COMPARTMENTS [148] which includes the data from UniProtKB [152] and HPA [141] with the updates from their recent publication Thul et al. [28].

tissue-specific SCL dataset So far, there is not yet a tissue-specific SCL dataset in public databases. Therefore, the collection of ground truth data and the corresponding evaluation are a challenge of this study. However, HPA provides the cell line information in which the SCLs of protein were detected in the cellular atlas [28]. As mentioned, cell lines were derived from corresponding tissue types. These cell line specific SCL data are expected to reflect some aspects of tissue-specific characters. The SCL data of human proteins were detected in different cell lines, and were validated by antibodies and IF microscopic images. It covers

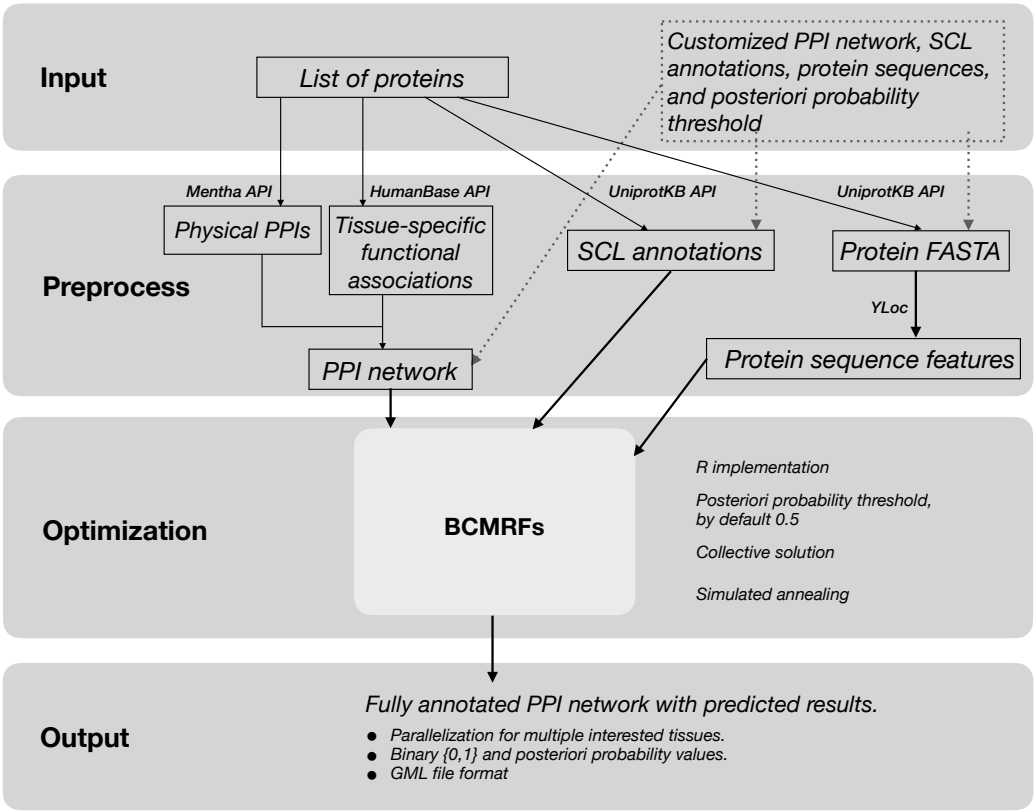


Fig. 5.2 The method is implemented in R and divided into the preprocessing and the optimization part. The R scripts for the preprocessing compute Furthermore, the user can also customize the PPIN, the SCL annotation, the protein sequences according to their research of interest. All the dataset are then preprocessed to match further computing. User can also improve precision or recall by setting the cutoff value for the posteriori probability value. The program is parallelized if multiple tissues are selected. As output GML formated files are provided for further visualization and modeling.

Table 5.1 Mapping table from cell lines to tissues.

CellLine	Tissue
SH-SY5Y	Brain
U-251 MG	Brain
AF22	Brain
A-431	Skin
HaCaT	Skin
SK-MEL-30	Skin
U-2 OS	Bone
RT4	Bladder
HeLa	Cervix
SiHa	Cervix
CACO-2	Colon
A549	Lung
HEK 293	Kidney
Hep G2	Liver

9841 human proteins in thirty SCLs for eighteen cell lines. All the SCLs of human proteins are assigned with a fluorescent intensity of 'Strong', 'Moderate' and 'Weak'. To increase the quality of the dataset, the 'Weak' intensity data were not considered. The cell lines and the detected SCLs data in these cell lines are mapped and grouped into corresponding tissues, see Table 5.1.

Protein-protein interaction network datasets

Physical protein-protein interaction network of human protein The generic physical PPI data were collected following the same criteria as described in Chapter 4. The data were retrieved from the interactome browser Mentha (version 2017.06) which limits itself to direct physical PPIs curated by members of IMEx. Each PPI is associated with a reliability score which is calculated based on the evidence such as experimental method, the size of experiments and relevant literature [49]. The PPIs which have a reliability score less than 0.25 were removed.

Tissue-specific functional associations Previously, Greene et al. [10] have studied human tissue-specific networks for understanding the multicellular function and disease of the human protein. They generated the genome-wide functional interaction networks for 144 human tissues and cell lines using a data-driven Bayesian methodology that integrates thousands of experiments spanning various tissue and disease states using both gene expression data

and ontology annotations. These networks were constructed using functional networks from genome-scale data by performing a tissue-specific Bayesian integration based on gene expression data of various types of human tissues and cell lines [10]. They estimated the probability of tissue-specific functional interaction between all pairs of genes. Only the top scored interactions were used in our study. The tissue-specific functional associations were downloaded via their HumanBase web services RESTful API `http://hb.flatironinstitute.org/api/`.

5.1.4 Performance measures

In the human protein SCL dataset, a protein might be associated with a set of SCL labels. Such a multi-label dataset is often imbalanced, meaning that some of the labels are very frequent whereas others are quite rare. The level of imbalance of a determinate label can be measured by the imbalance ratio, IRLbl Equation (2.4). The imbalance level of our multi-label dataset of human proteins is represented by Mean Imbalance Rate (MeanIR) among all labels, see Equation (2.5).

To evaluate the prediction performance of our method, the SCL labels of one-third of the proteins from the protein pool, which have SCLs annotation, were masked and treated as unknown proteins. We keep the labelings of the remaining proteins for training which are typically called 'seed'. The predicted labels of these masked proteins are later used for performance evaluation.

The traditional performance measures are difficult to apply for multiple SCL prediction. To better reflect the multi-label capabilities of classifiers, multi-label classification requires more sophisticated performance metrics than single-label classification. Popular multi-label measures include Accuracy, Precision, Recall, and F1-score and Hamming loss, see Section 4.2.2. Except for Hamming Loss, for all the remaining performance measures, the higher the measures, the better the prediction performance.

5.2 Results

5.2.1 Statistics of the tissue-specific physical PPINs

At first, the analysis of the characters of tissue-specific physical PPINs were conducted. Figure 5.3a demonstrates that both the size of the tissue-specific PPINs and the average degree of networks dramatically decrease to less than the half of these values of the generic physical PPI networks. It confirms that a big amount of PPIs are mutually exclusive across the listed tissues.

Each physical PPI is associated with a reliability score which takes the evidences such as experimental method, size of experiments and relevant literature into account [49]. Figure 5.3b shows the average of the reliability score of the networks. The reliability scores of all the tissue-specific PPINs are higher than these of the generic PPI network. That means the filtering process can remove spurious PPIs and improve the quality of the networks which is very important to have better SCL prediction.

Moreover, by computing the overlap of PPIs occurring in each tissue-specific PPIN, the similarity of different tissue-specific PPINs can be investigated (Figure 5.3c). Some tissue-specific PPINs have very high similarity, such as Liver, Kidney and Lung tissue. It indicates that tissues with similar functions tend to contain more common interactions. By contrast, the Bladder and Cervix tissue show the least similarities with all the other tissues, which indicates that they have high tissue specificity. Figure 5.3d shows that, in general, there is a relatively high percentage of common interactions across different tissues (above 67% overall), indicating different tissues share similar interactions or working mechanism despite their different functions. These interactions might occur among the proteins encoded by the housekeeping genes which maintenance of basic cellular function for cellular survival.

5.2.2 Statistics of the tissue-specific SCLs

When take a closer look at each SCL class of each tissue, the imbalance level of SCL differs from one tissue to another . Bladder, Cervix, and Kidney are more imbalanced in protein SCLs than the rest of listed tissues (Table 5.2). Table 5.3 shows the details of the distribution of proteins SCLs in nine tissues. The majority of proteins are localizing in the Nucleus for all the tissues. Proteins which access to the Nucleus are often highly regulated and controls critical steps in development, stress response, and general cell signaling [153]. The next large number of proteins was identified in the Cytosol, Vesicle and Plasma membrane which are the key participants of the secretory pathway in the cell.

The comparison of the tissue-specific SCL of each protein with the generic SCL data for the listed tissues shows that the ER-proteome and Plasma membraneproteome vary dramatically among the tissues. Next, the Golgi apparatus proteome of the bone tissue and the Cytoskeleton proteome of the brain tissue differ a lot. On the other hand, for all the tissues, the Nucleus- and Cytosol-locating proteins are similar, see Figure 5.4b.

The human proteome shows a variance of their cell-to-cell spatial distribution, as well of their cell line-dependent location with different localization in the three cell lines tested [28]. The tissue specificity of protein SCL can be investigated by computing the agreement of SCLs of each protein occurring in each tissue (Figure 5.4a). Skin and Brain show the highest similarity (59%), while Bladder and Brain show the lowest similarity (52%). The tissue pairs

Table 5.2 The imbalance level of SCL dataset across tissues.

	Cytoskeleton	Cytosol	ER	GA	Mito	Nucleus	PM	Vesicle	meanIR
Brain	6.825	1.378	22.612	8.634	7.436	1.000	5.310	5.842	7.380
Bone	6.407	1.570	17.861	10.419	6.154	1.000	5.199	4.254	6.608
Skin	6.993	1.458	23.758	8.070	6.770	1.000	4.665	4.805	7.190
Bladder	8.645	2.653	24.364	7.882	13.400	1.000	5.469	3.671	8.386
Lung	8.220	2.035	20.550	8.220	6.524	1.000	5.408	3.841	6.975
Colon	12.655	3.745	22.938	10.794	7.058	1.000	5.169	3.745	8.388
Liver	14.318	2.582	9.545	5.000	6.562	1.000	7.326	4.257	6.324
Cervix	7.511	2.651	19.882	7.953	9.941	1.000	6.377	4.173	7.436
Kidney	10.815	2.454	21.630	8.588	7.892	1.000	9.270	4.908	8.319

Table 5.3 The distribution of protein SCL across tissues.

	Number	Cytoskeleton	Cytosol	ER	GA	Mito	Nucleus	PM	Vesicle	Sum
Brain	5632	487	2413	147	385	447	3324	626	569	8398
Bone	8873	800	3264	287	492	833	5126	986	1205	12993
Skin	6456	547	2623	161	474	565	3825	820	796	9811
Bladder	431	31	101	11	34	20	268	49	73	587
Lung	668	50	202	20	50	63	411	76	107	979
Colon	569	29	98	16	34	52	367	71	98	765
Liver	541	22	122	33	63	48	315	43	74	720
Cervix	1037	90	255	34	85	68	676	106	162	1476
Kidney	920	54	238	27	68	74	584	63	119	1227
Sum	25127(10281*)	2110	9316	736	1685	2170	14896	2840	3203	36956

* without repetition across tissues.
Aberrations: ER (Endoplasmic Reticulum), Mito (Mitochondrion), GA (Golgi Apparatus), PM (Plasma Membrane)

Colon-Liver, Bone-Brain and Bone-Skin share more common spatial distribution of proteins in the cell than the other tissues. When we compare the agreement of SCLs across all the nine human tissues, it shows a low similarity of protein SCL with the average similarity of 8.45%. A low similarity indicates typically a high tissue specificity of protein SCLs. One explanation would be that the number of proteins which have SCLs across tissues varies from 431 to 8873 (Table 5.3). The overlap of proteins itself is rather low. The most of the proteins whose SCLs are detected are expressed in Bone, Skin and Brain tissue. However, within these highly annotated tissues, the level of tissue specificity on protein SCsL is still rather high. Afterwards, the tissue-specific SCL dataset was compared against the generic SCL dataset of the human protein (Figure 5.4b). The result shows that tissue specificity exists for the proteomes in most of the SCLs. The Cytoskeleton proteome, Plasma membrane proteome, and ER proteome are significantly different across tissues. The Cytosol proteome shows its specificity in Colon and Bladder tissue. On the other hand, there is nearly no difference of the Nucleus-proteome across nine tissues.

5.2.3 The impact of the noisy tissue-specific functional associations on tissue-specific SCL prediction

The proposed prediction method for tissue-specific SCL highly relies on the tissue-specific functional associations. However, the tissue-specific functional association data which are used in this study were generated by a Bayesian approach on tissue-specific gene expression data [10]. In this section, we analyze the impact of the noise of these predicted data effecting the prediction of the tissue-specific SCL.

To address this question, we performed the experiments based on the tissue-specific PPINs with false interactions. To generate such PPINs, we employed the published specific gold standard tissue-specific functional associations from Greene et al. [10] for each tissue including positive and negative functional associations. The gold standard dataset includes

- True tissue-specific functional associations: the positive functional edges between genes specifically co-expressed in the tissue.
- False tissue-specific functional associations include three scenarios. To construct comparable networks across tissues, we used a negative set composed of equal proportions of edges from the three scenarios.
  - the positive functional edges between a gene expressed in the tissue and another specifically expressed in an unrelated tissue.

- the negative functional edges between genes specifically co-expressed in the tissue.
- the negative functional edges between one gene expressed in the tissue and another specifically expressed in an unrelated tissue.

The number of the tissue-specific functional associations in the golden standard dataset vary across tissues according to the specificity of the tissue, the depth of the study, and how well curated from literature the data are [10]. The size of the association set differs across tissues due to the incompleteness of the tissue-specific SCL data from HPA. Because of the size of golden standard dataset, we limited the percentage of noise to 50% to have enough number of proteins to evaluate the performance. Following the above criterion, the tissue-specific PPINs with noise across tissues were constructed. We performed the tissue-specific SCL predictions using those protein-protein interaction network and evaluated the results against the tissue-specific SCL ground truth dataset.

The resulting changes of prediction performance along increasing rate of noise are shown in Figure 5.5. In general, the proposed prediction method is sensitive to the topology of the tissue-specific PPINs. We assume the lower the quality of PPINs, the less accurate the prediction. This is confirmed from the significant decline of performance of Lung (0.8 to 0.6) and Liver (0.75 to 0.6), and slight decrease of Skin (0.75 to 0.7) and Colon (0.65 to 0.57). However, the curves of Kidney, Bone, and Brain appear slight ups and downs. The performances with and without noise PPI remain rather still. On average, the performances tend to decrease while increasing the noise in the PPIN. However, the average change (0.77 to 0.68) is not strikingly large, because the tissue-specific PPI networks were constructed from the filtered functional associations by physical protein-protein interaction. Thus, a number of the noise PPI were already filtered out by physical PPIs.

To ensure the quality of the tissue-specific functional associations and reduce the impact of the noise, we used two ways to extract meaningful protein-protein interaction.

1. Only the associations with evidence supporting a tissue-specific functional interaction are used in SCL prediction.

2. The posterior probability of tissue-specific functional associations was used additionally to weight the edges in the tissue-specific PPINs. The edge weight is the product of the reliability score of physical interaction and the posterior probability of tissue-specific functional associations. Thus, the less reliable associations which have low probability values would have less impact on the model.

5.2.4 Genome-wide tissue-specific SCLs prediction

As described in Section 5.1 and illustrated in Figure 5.1, to predict tissue-specific SCL, the BCMRFs algorithm was applied to tissue-specific physical PPINs to predict human proteins on nine tissues (including Brain, Bone, Skin, Bladder, Lung, Colon, Liver, Cervix and Kidney) focusing on eight major SCLs (including Cytoskeleton, Cytosol, Endoplasmic reticulum, Golgi apparatus, Mitochondrion, Nucleus, Plasma membrane and Vesicle). By comparing the predicted SCLs among each tissue, and with the generic SCLs, a landscape of dynamic changes in SCLs across tissues can be revealed.

To test the performance of the tissue-specific SCL predictions, an evaluation of the prediction results against the only available tissue-specific SCL which were generated based on the cell line-specific SCLs as ground truth was performed using multi-label classification metrics (see Section 5.1).

BCMRFs associates each protein in the PPIN with an estimated probability value for each SCL. Typically, the final SCL labels are generated by thresholding the probabilities. A threshold of 0.5 (logistic distribution) was used as a baseline for the evaluation. As it can be seen in Table 5.4 (values in brackets), the prediction performances of tissue-specific SCL are fair. The best performance is achieved on Kidney tissue (F1 score 0.718), while the weakest prediction is on Bladder (F1 score 0.616). Although it obtains a good recalls (from 0.825 to 0.898), the precisions are relatively low (from 0.539 to 0.655), which decreases the overall F1 scores.

This is not surprising since that generic SCL data was used as input and the tissue-specific PPINs share on average 50% PPIs. Moreover, the protein SCL datasets are highly imbalanced, especially for SCL such as ER, and a threshold of 0.5 is an arbitrary decision which cannot always produce the correct labeling result. Therefore, the optimal threshold were estimated for each SCL class according to the probabilities of the 'seed' proteins, which were used for determining the SCL labeling on the graph. Most of the thresholds increased to about 0.6, except for Nucleus which is about 0.4. Table 5.4 shows the evaluation results on the 'masked' proteins (see Section 5.1). As expected, the precision increased with the higher thresholds which increases the F1 scores overall.

It is important to remember that the prediction performances do not reflect the real strength of our approach. Because the quality of our ground truth datasets is limited. These ground truth data are generated based on detected protein SCL in cell lines. First of all, these cell lines are chosen for SCL experiments because of the easier manipulation procedure and the better results, especially for U2-OS cells [28]. In addition, some of the cell lines are cancerous cell lines, and some are immortalized cell lines. The tissue-specific functional associations were also generated based on various types of datasets using both tissues and

cell lines. Hence, the healthy and the diseased tissues were not distinguished in this study. In general, cell lines cannot represent all the tissue-specific features due to down-regulation of tissue-enriched genes [8].

5.2.5 Predictions for novel tissue-specific protein candidate validated by text mining

In total, 1863 proteins which show tissue specificity on SCL across nine listed tissues were identified. 1314 of these proteins had previously been found to show cell line dependency by Thul et al. [28].

The remaining 549 proteins which were newly found differentially localizing across nine tissues, were evaluated by text-mining. 243 of the candidate proteins and their SCLs in corresponding tissues are found in 724 publications in total. In the following, several illustrative examples which were validated by literature were shown. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was proven that it participates in nuclear events including transcription, RNA transport, DNA replication, and apoptosis. However, its over-expression and the increased enzymatic activity in proliferating cells, with preservation of its cytoplasmic localization, would occur in response to the elevated energy requirements of dividing hepatocytes [154]. Exocyst complex component SEC10 (Sec10) is a crucial component of the exocyst complex. It was previously proven that in renal tubular epithelial cells, Sec10 colocalizes with Cdc42 at the primary cilium which is bounded by an extrusion of the plasma membrane [155]. Synapsin-1 protein (SYN1) together with CA1 and CA3 synaptosomes are co-localized on the same synaptic vesicles in mossy fiber nerve terminals of the hippocampus [156]. Ubiquitin-like protein 5 (UBL5) was found localized in the nucleus, partially associates with Cajal bodies in human embryonic kidney cells [157]. Furthermore, it was found that Nuclear factor erythroid 2-related factor 2 (Nrf2) was expressed both in the cytoplasm and nuclear of glomeruli and tubules [158]. These results demonstrate that the tissue specificity of protein SCL is essential to carry out their specific functions across tissues.

5.3 Summary

This chapter described an approach that allows predicting protein SCL on tissue specificity through the use of tissue-specific functional associations. To the best knowledge, this is the first computational approach addressed the tissue-specific SCL of a large number of human proteins. It is an extension study of previously developed multi-SCL prediction

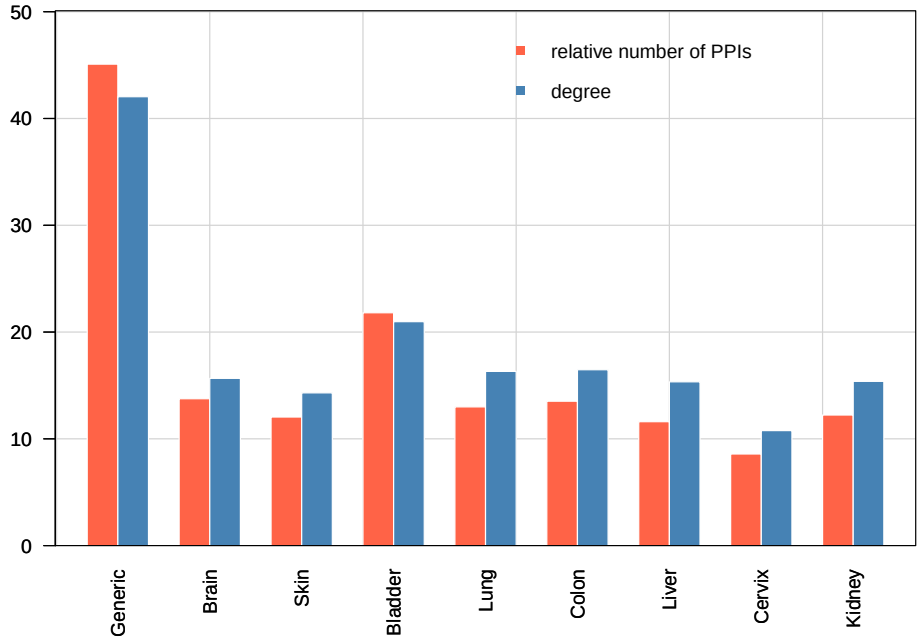
Table 5.4 tissue-specific multi-SCL prediction performance.

	Average precision	F1 score	Hamming loss	Precision	Recall
Brain	0.869 (0.713)	0.71 (0.669)	0.157 (0.188)	0.665 (0.592)	0.868 (0.896)
Bone	0.87 (0.699)	0.707 (0.655)	0.157 (0.191)	0.664 (0.575)	0.867 (0.893)
Skin	0.87 (0.698)	0.708 (0.657)	0.158 (0.195)	0.667 (0.575)	0.862 (0.894)
Bladder	0.822 (0.708)	0.653 (0.616)	0.188 (0.202)	0.602 (0.539)	0.822 (0.825)
Lung	0.865 (0.704)	0.696 (0.66)	0.175 (0.188)	0.65 (0.585)	0.865 (0.893)
Colon	0.886 (0.696)	0.733 (0.678)	0.137 (0.164)	0.668 (0.605)	0.915 (0.898)
Liver	0.869 (0.764)	0.721 (0.691)	0.138 (0.15)	0.691 (0.65)	0.853 (0.859)
Cervix	0.852 (0.689)	0.705 (0.642)	0.153 (0.181)	0.661 (0.58)	0.848 (0.845)
Kidney	0.891 (0.737)	0.736 (0.718)	0.131 (0.136)	0.692 (0.655)	0.885 (0.897)

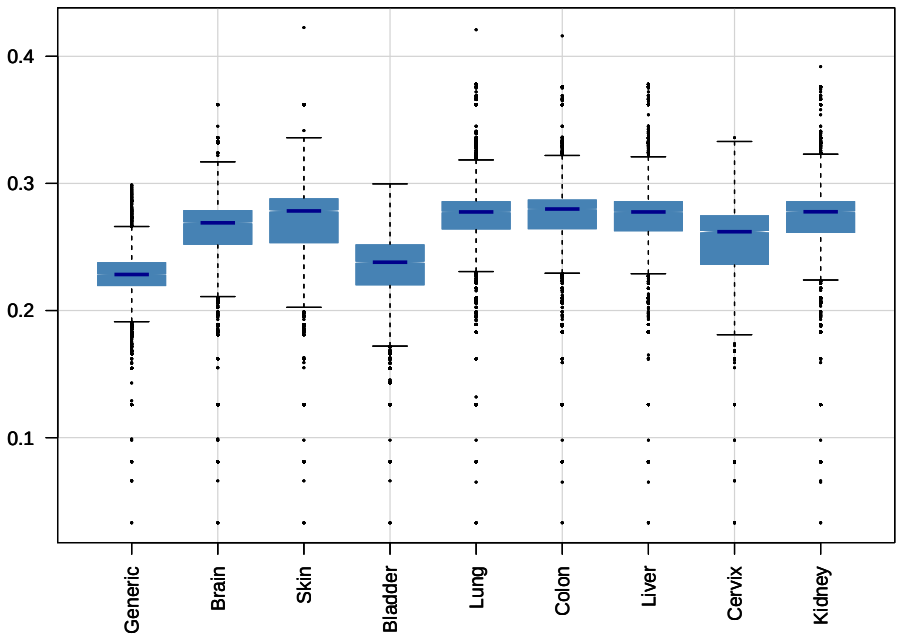
(*): The prediction performances using 0.5 as threshold.

method BCMRFs on tissue-specific SCL prediction. In this chapter, both tissue-specific physical PPINs and tissue-specific SCLs ground truth data were analyzed and proved their particular characters across different tissues. The BCMRFs algorithm was applied on physical PPINs filtered by tissue-specific functional associations for nine types of tissue focusing on eight high-level SCLs. The evaluated results demonstrate the strength of our approach on predicting tissue-specific SCLs. In total, 1314 proteins which were previously proven cell line dependent on SCL level were successfully identified. Furthermore, 549 novel tissue-specific localized candidate proteins were predicted and some of them were validated via text-mining. These candidates should be verified experimentally in the future.

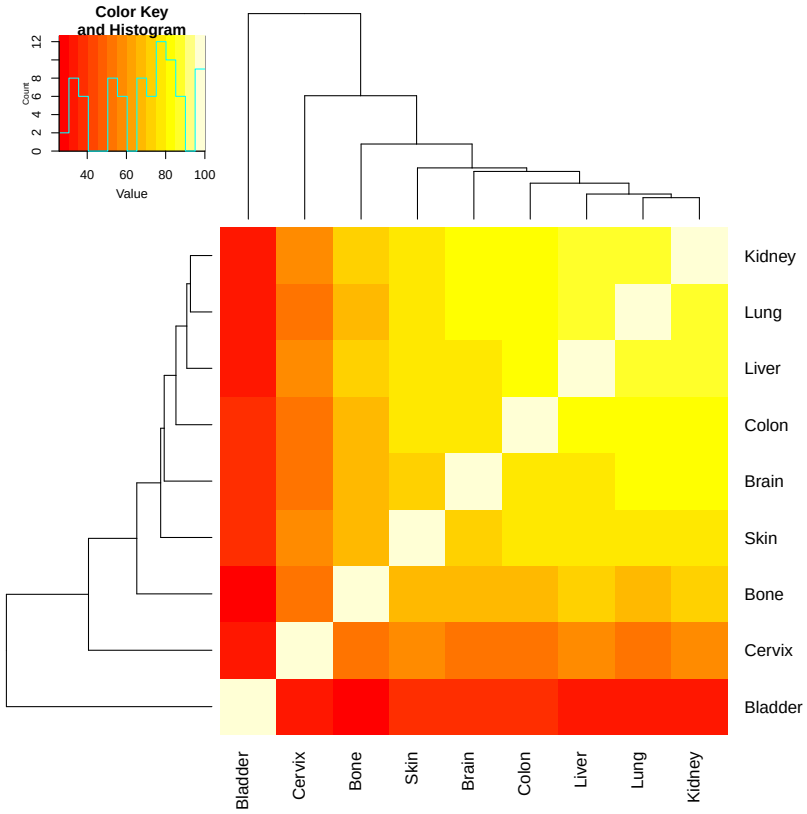
Knowing the tissue specificity of protein on subcellular level would provide the insights for identifying the changing functional roles of genes across tissues and illuminate relationships among diseases. There is clear gap of the research area given the intensive research on tissue-specificity proteome and SCL-specific proteome, and the lack of ground truth data of tissue-specific protein SCLs. Although some efforts have been made to collect such data by the Encyclopedia of DNA Elements project (ENCODE) [80], the quantity of data is very limited to several SCL for human proteins. It is highly recommend that researchers specify the tissues/cell lines where their experimentally detected protein SCL and PPIs occurred in, and this information should be differentiated in the public databases such as UniProtKB, HPDB, which would make it easier to model and understand context-related phenotypes. On the other hand, an alternative would be using text-mining approach to collect the data from diverse literature.



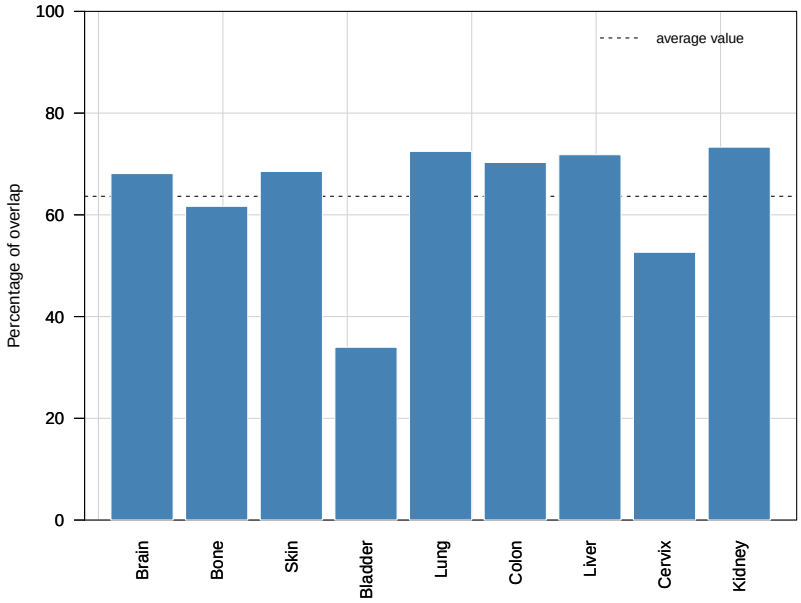
(a) The mean degree value of the PPI networks.



(b) Average reliability scores of the PPI networks.

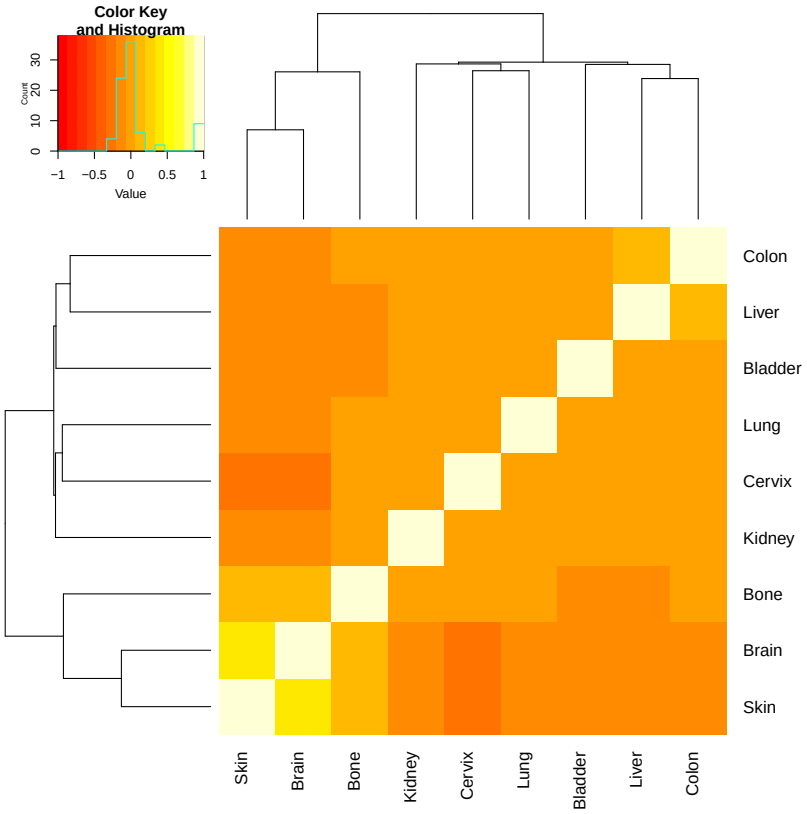


(c) Similarity of tissue-specific PPINs.

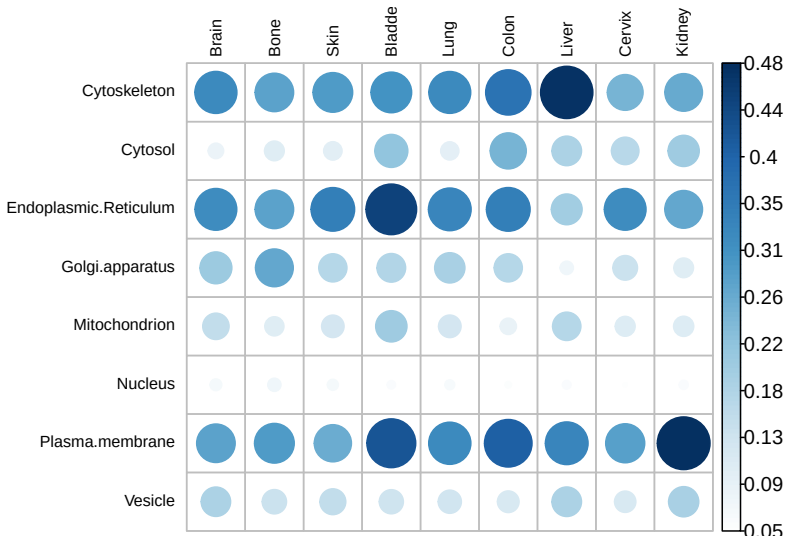


(d) The overlap percentage of tissue-specific PPINs.

Fig. 5.3 The property of the tissue-specific physical PPINs. (a) Comparison of generic PPI network and tissue-specific PPINs on the size of the network and the degree of the network. (b) Comparison of the generic PPI network and tissue-specific PPINs on the reliability score of physical PPIs. (c) The heatmap for the similarity among all the tissue-specific PPINs. The brighter the color, the more similar of a pair of tissues. (d) The percentage of the common PPIs of all tissues.



(a) The heatmap of TP-SCLs of human proteins.



(b) The changes of SCL on each tissue.

Fig. 5.4 Comparison of protein SCLs across tissues. (a) The similarity of protein SCL datasets among nine tissues. The brighter the color, the more similar between two tissues. (b) The changes of protein SCL across 9 tissues in comparison with generic protein SCL dataset. The color and the size of circle positively correlate to the difference and specificity of the SCL.

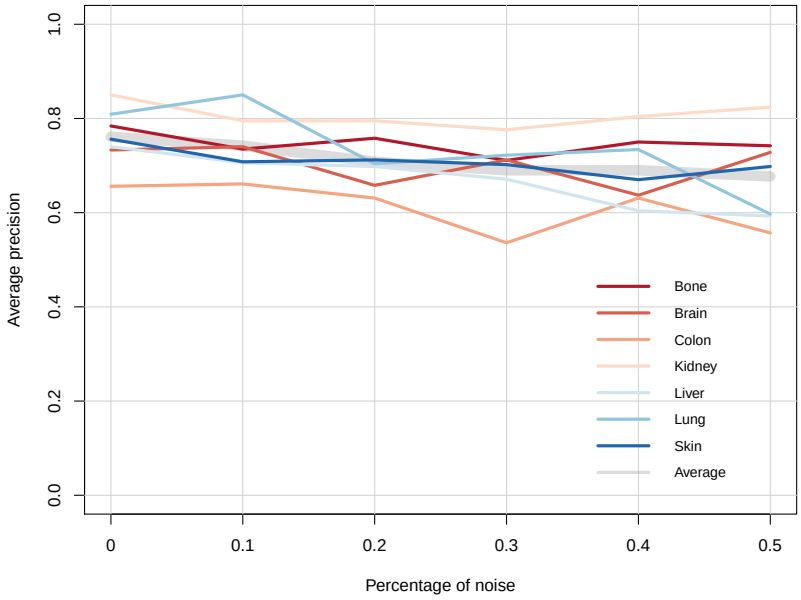


Fig. 5.5 Impact of tissue-specific functional association on performance.

Chapter 6

Tissue-specific SCL Data Curation using Text mining

Knowing the tissue specificity of protein on the subcellular level can provide the insights for identifying the protein functions across tissues, illuminate the fundamental mechanisms of the human cells. In the previous chapter, we discussed the necessity of knowing tissue-specific SCL of protein for human biology. Facing the fact that there is not yet any protein tissue-specific SCL dataset available in public databases, the alternative to having such data collection is to extract the data from the scientific literature which is the challenge of the research in this chapter. The biomedical literature is the key distribution channel for novel findings and hypotheses from biochemical and biomedical research worldwide.

The automated literature analysis, text mining approach, is now frequently a part of sophisticated biomedical research which retrieves relevant information and identifies the connections between pieces of knowledge from numerous publications. Such automated analysis of literature complements the reading of scientific literature by individual researchers as it allows rapid access to information from a large number of documents and may increase the reproducibility of literature research.

Text mining approaches have been successfully applied to the identification of molecular causes of diseases [159], PPIs [160, 161], the interactions of small molecules and proteins, the influences of genetic variation on drug responses [162] and many other types of research. This chapter describes how to apply text mining approaches to extract the information of protein tissue-specific SCLs. The primary tasks include the information extraction of proteins, the expressed tissue, and their spatial distribution in the cell in the tissue.

Before introducing the approach, it is necessary to mention the related works. Previously, some contribution has been made to the extraction of protein SCL from text, such as Textpresso [163] and COMPARTMENT [148]. The tissue-specificity of genes and the encoded

proteins have been addressed using text mining approaches as well [164]. However, the above data curation on protein SCL and tissue-specificity of protein were studied individually as the paired association between protein and SCL, and protein and tissue. Therefore, the information extraction on the triple association of tissue, protein, and SCL remains as an open question in systems biology and a challenge for text mining.

To date, there are only a few studies which have addressed the information extraction on the triple association. In 2016, Mahmood et al. [165] and Singhal et al. [166] both published the works on the extraction of triple association of disease, gene, and mutation using the text mining system. The former work extracts the information of mutation, paired association <mutation, gene> and <mutation, disease> on three steps, and later link these three elements to create the triple association of <disease, gene, mutation>. In addition to the co-occurrence-based approach, they use sentence structure together with other textual features to increase the confidence of the information extraction [165]. Differently than Mahmood et al. [165], [166] performed the information extraction on gene and the paired association <mutation, gene> followed by gene sequence validation to identify an exact gene match for the mutation [166].

In general, the underlying mechanisms used for hypothesis generation and knowledge discovery range from basic co-occurrence techniques to complex machine-learning algorithms for identifying meaningful relationships among the extracted scientific facts. Co-occurrence analysis identifies named entities that are mentioned together in a portion of text, such as a sentence, a paragraph, a section or a whole document. Co-occurrences are then analyzed using statistical approaches and methods from information theory to identify important novel related terms. Co-occurrence can be indicative of a biological relationship between the identified entities and therefore leads to a novel hypothesis that can be examined in the future. Both of the above studies have used the co-occurrence approach as the baseline method.

In contrast to the complicated relationship between gene and mutation and the mentions in the text, in the case of the triple relationship of tissue, protein, and SCL, it is rather straightforward. In addition to the previous work which applies co-occurrence approach to extracting <protein, SCL> and <protein, tissue>, we are confident to succeed in the extraction of triple association <tissue, protein, SCL> using a co-occurrence based text mining approach. The general idea is to use a scoring system which takes into account the co-occurrences within sentences and whole documents. Later we combine them through an optimized weighting scheme to distinguish the true positive triplet from the false positives.

6.1 Methods

Figure 6.1 shows the schema of the overall architecture of the system is shown. The system can be divided into eight main tasks, namely from A to G. Task A retrieves the abstracts of the relevant publications fetched from MEDLINE abstract which contains gene/protein, tissue, and SCL in the text. Task B applies some primary text preprocessing on these abstracts, such as tokenization. Then, the various types of mention of tissue, gene/protein and SCL terms in the text are recognized and linked to the appropriate, unique identifiers, such as identifiers as Entrez gene ID in NCBI, the BRENDA Tissue Ontology (BTO) identifier in BRANDA and GO identifier (Task C and D). In Task E, information extraction, in addition to the co-occurrence of above three types of mentions, we applied a scoring function for identifying the important relevant associations. The extraction of the information such as the position of above mentions in the text and across all the abstracts was conducted as well. The scoring system was implemented in Python 3 [167]. The following subsections describe the details of each task.

6.1.1 A. Retrieving relevant abstracts

A set of abstracts that were potentially useful for identifying tissue, gene/protein, SCL was retrieved from MEDLINE via PubMed search engine. The MEDLINE index of abstracts contains Medical Subject Headings (MeSH) which are humanly curated annotations for a controlled vocabulary of biomedical concepts. Thus, we can restrict the retrieved abstract set to a specific species, tissue and SCL of interest using the ‘E-utilities’ programming interface [168]. For example, we can submit the query

"kidney"[MeSH Terms] OR "kidney"[All Fields]) AND "human"[MeSH Terms]

to access all the abstracts which are related to human kidney tissue and do not necessarily have to contain the exact search term.

6.1.2 B. Text preprocessing

Before any specific text processing was initiated, the identification of the individual sentences of documents was performed. The following step is to ’tokenize’ the document which is for identifying the constituents of the text (which are called ’tokens’), such as single words, numbers or punctuation. Consecutive tokens can be combined to form named entities of tissue, gene/protein, and SCL.

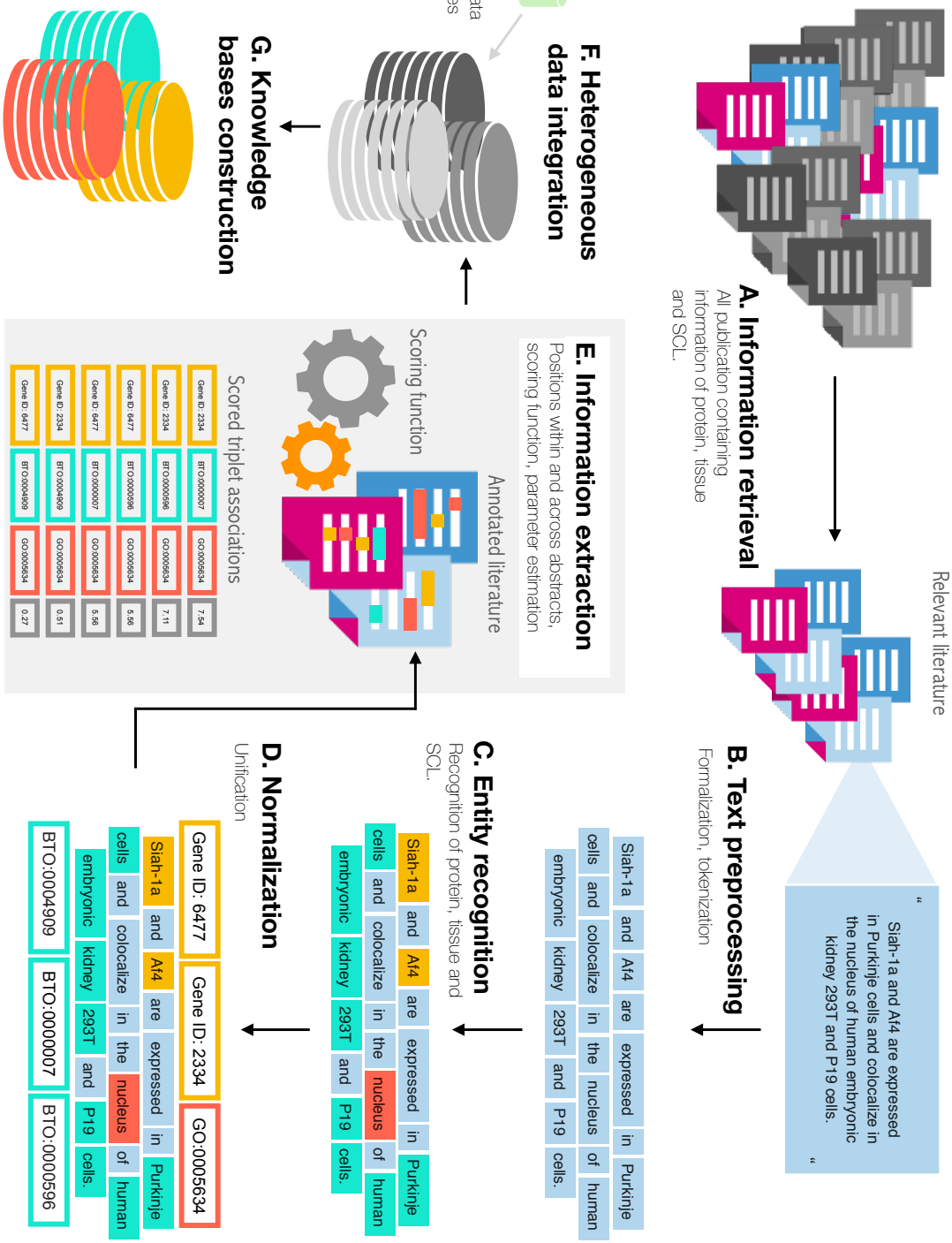


Fig. 6.1 Schematic diagram of proposed text mining system.

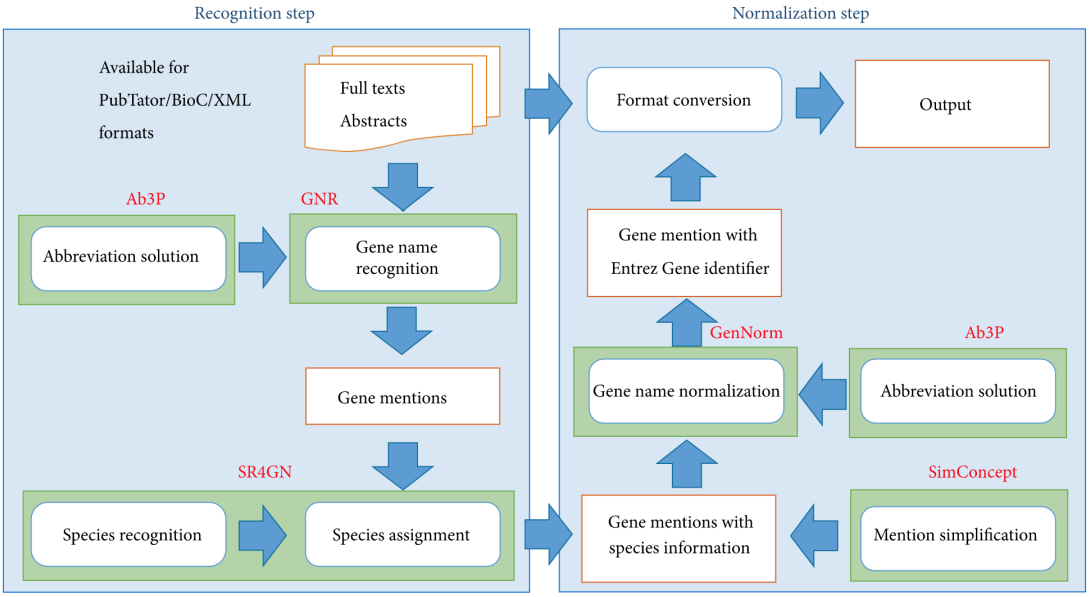


Fig. 6.2 The overview of GNormPlus method from Wei et al. [170]

6.1.3 C. Mamed entity recognition

The follow-up step is to identify the named-entities of tissue, protein, and SCL in the text. The identification of the correct boundaries of composed terms and the disambiguation of terms is a complicated process. In this task, different approaches to identify the three types of named entities were employed.

Identification of protein mentions

The NER of gene mention in the text is always an essential part of solving various hypotheses in biology using text mining. For that matter, many algorithms and tools have been developed and successfully applied in natural language processing (NLP) research. In this task, we included the state-of-art text mining tool PubTator [169] which can tag various biological entities. PubTator employs the high performance (F1-score 86.7%) entity recognition tools including GNormPlus [170] for the identification of gene/proteins from a given text, see Figure 6.2. The gene/protein entity mentions are recognized using a conditional random field based module in combination with the species recognition module SR4GN [171]. Next, the gene/protein mention are normalized using their previously developed tool GenNorm [172] in combination with the composite mention simplification tool - SimConcept [173] and an abbreviation resolution tool Ab3P [174].

Identification of SCL and tissue mentions

Unlike the NER process for gene/protein, there is no ready-to-use NER tool available for neither SCL nor tissue, yet. Therefore, for these two types of entity, we apply the tissue and SCL detector developed in-house. The detector uses a dictionary-based approach which relays on matching a dictionary of names against the text. For this purpose, the carefully curated dictionaries for tissue and SCL are essential for the good performance of mining the important associations.

Dictionary construction Ontologies which are used in life science represent classification systems that provide a controlled vocabulary for a biological or biomedical knowledge domain. An important pioneering effort in the field of biological ontologies, probably being the most widely used, is theGO project (`http://www.geneontology.org`) that is a collaborative effort to develop and use ontologies to support biologically meaningful annotation of genes and their products. GO contains a set of terms for describing the activity and actions of gene products. Each of these activities is executed in a SCC or a location outside in the vicinity of a cell. In order to capture this context, the GO includes a sub-ontology called the cellular component ontology (GO-CCO). The GO-CCO describes subcellular structures and macromolecular complexes. Moreover, BTO (`http://www.BTO.brenda-enzymes.org`) represents a comprehensive structured encyclopedia of tissue terms which is a connection between the enzyme data collection of the BRENDA enzyme database and a structured network of source tissues and cell types. BTOcontains different anatomical structures, tissues, cell types and cell lines, classified under generic categories corresponding to the rules and formats of the Gene Ontology Consortium and organized as a directed acyclic graph (DAG) [175]. Likewise, Uber-anatomy ontology (Uberon) (`https://uberon.github.io/about.html`) is an anatomical ontology that represents body parts, organs and tissues in a variety of animal species, with a focus on vertebrates [176].

The dictionaries were automatically generated based on the annotations and the synonyms contained in the cellular component terms from GO for SCL, and terms from BTO and Uberon for tissue respectively. To reach better recall, the variants of term which include

- the conversion of the upper case and lower case of the first letter of the term, except the abbreviations.
- the pluralization of terms.
- remove highly repeated common patterns across all terms, unless this would cause ambiguity, such as 'inner', 'channel'.

- removal the brackets, parentheses, and hyphens.

were automatically generated as well.

Afterwards, the dictionaries are manually curated, such as to eliminate synonyms that give rise to many false positives. The ambiguous terms which occur in other categories were removed from the dictionaries. For example, 'spindle' (GO:0005819) is the array of microtubules and associated molecules that forms between opposite poles of a eukaryotic cell during mitosis or meiosis and serves to move the duplicated chromosomes apart. Whereas 'spindle' in spindle cell (BTO:0003651) represent the fusiform cell, such as those in the deeper layers of the cerebral cortex.

6.1.4 D. Term normalization

The term normalization, which is also known as Named-entity Linking, process contributes to the integration of literature with data contained in biomedical resources. Dictionary-based methods have the crucial advantage of being able to normalize names. Each annotation terms from GO, BTO and Uberon corresponds to a unique identifier in the ontology. While constructing dictionaries, the identifiers are already included which facilitates the term normalization step.

The texts were subsequently parsed into individual sentences, tokenized words to match against the dictionary. Afterward, by string matching of the dictionaries against the text, the entities of SCL and tissue is, thus, recognized. In dictionary-based approach, the SCL entities and tissue entities which are mentioned in the text are fitted to the best match from the dictionary resources and then immediately linked to database entries of GO and BTO, respectively.

6.1.5 E. Extraction and scoring of tissue-protein-SCL associations

The next task is to extract information of tissue, gene/protein, and SCL and identify their triple relations. The identification of relations is a more complicated task and can be accomplished with different methods. One popular NLP approach is to use a grammar to parse the syntax of each sentence. In this study, we used a statistical co-occurrence based analysis, which determines triplets of named entities that are mentioned together in a portion of text.

Previously, Franceschini et al. [160] successfully applied a scoring schema for determining associations between paired proteins derived from their co-occurrence. They found the co-occurrence based method is much more flexible and gave better recall in their study. This scoring schema takes into account co-occurrences within sentences, within paragraphs, and

within whole documents and combines them through an optimized weighting scheme. Later, this schema was implemented to determine disease–gene associations [177] and protein SCL [148].

Motivated by the previous success on the extraction of paired associations, for the first time, the scoring schema was modified and applied to solve the triple association extraction problem in this thesis. The detailed formulation is explained in the following subsection.

Co-occurrence scoring function

An important feature of the co-occurrence scoring scheme is that it simultaneously takes into account co-occurrences at the level of abstracts as well as individual sentences. As shown in Algorithm 4, to score the co-occurring tissue, gene/protein, and SCL triplet, we first calculate the weighted count $C(T,P,L)$ of the triplets appearing over the n abstracts in the text corpus. $C(T,P,L)$ is defined in below:

$$C(T,P,L) = \sum_{k=1}^{n} w_s \delta_{sk}(T,P,L) + w_a \delta_{ak}(T,P,L) \tag{6.1}$$

where n is the number of abstracts, w_s and w_a are the weights for co-occurrence within the same sentence, and within the same abstract, respectively. If P ,T and L are mentioned together in a sentence or in abstract k, the delta functions $\delta_{ak}(T,P,L)$, $\delta_{pk}(T,P,L)$ and $\delta_{sk}(T,P,L)$ are 1, and 0 otherwise. Thus, an abstract that mentions P ,T and L in the same sentence will give a score contribution of $w_s + w_a$, whereas an abstract that mentions them in different sentences will give a score contribution of w_a only. The scoring function is therefore defined as

$$S(T,P,L) = C(T,P,L)^{\alpha} CoP^{1-\alpha} \tag{6.2}$$

where α is the partition parameter.

$$CoP = \frac{Pr^*(T,P,L)}{Pr^*(P) \cdot Pr^*(T) \cdot Pr^*(L)} \tag{6.3}$$

with

$$\begin{aligned} Pr^*(T,P,L) &= \frac{C(T,P,L)}{C(\star,\star,\star)} \\ Pr^*(P) &= \frac{C(P,\star,\star)}{C(\star,\star,\star)} \\ Pr^*(T) &= \frac{C(\star,T,\star)}{C(\star,\star,\star)} \\ Pr^*(L) &= \frac{C(\star,\star,L)}{C(\star,\star,\star)} \end{aligned} \tag{6.4}$$

$Pr^*(T,P,L)$, $Pr^*(P)$, $Pr^*(T)$, $Pr^*(L)$ are the observed probability of triplet, protein, tissue and SCL based on weighted count, respectively.

Therefore, the co-occurrence score of triple association $S(T,P,L)$ can be written as follows :

$$S(T,P,L) = C(T,P,L)^{\alpha} \left(\frac{C(T,P,L)C(\star,\star,\star)^2}{C(P,\star,\star) \cdot C(\star,T,\star) \cdot C(\star,\star,L)} \right)^{1-\alpha} \tag{6.5}$$

where $C(P,\star,\star)$ is the sums over all the possible tissue and SCL pairs associated with protein P. Likewise, $C(\star,T,\star)$ is the sum over all the possible protein and SCL pairs associated with tissue T, and $C(\star,\star,L)$ is the sum over all the possible protein and tissue pairs associated with SCL L. Moreover, $C(\star,\star,\star)$ is the sum over all possible triplet of proteins,tissues and SCL. The parameters were optimized on the benchmark set which are shown in the result section in this chapter.

Normalized Z-scores

The score $S(T,P,L)$ depends on the number of triplets $< T,P,L >$ identified in the abstract pools, which changes as the number of abstracts grows, see Equation (6.1). To get a more robust measure, therefore, the scores are converted into the normalized z-scores $Z(T,P,L)$ relative to a background distribution.

Assume that the empirically observed score distribution is a mixture of lower-scoring random background and the higher-scoring true signal. We model the background distribution as a Gaussian and estimate its mean as the mode of the mixture distribution. We empirically observed that the 40^{th} percentile, in this case, coincides with the mode which is accord with the observation from Binder et al.'s work. The variance of the background is estimated from the difference between the 20^{th} and the $40th$ percentiles, -1.282 and -0.842 respectively.

$$z = (x - \mu)/\sigma \tag{6.6}$$

that μ is the mean of the score population, σ is the standard deviation of the score population.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N-1}} \tag{6.7}$$

with

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{6.8}$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed score values of present triplets.

The absolute value of z represents the distance between the raw score and the population mean in units of the standard deviation. z is negative when the raw score is below the mean, positive when above.

Algorithm 4: Triplet association extraction

Input: Abstracts tagged with three entities, T, P, L with their position in the text
Output: Extracted <T,P,L> triplets associated with the co-occurrence Z-scores

for *each abstract* **do**
  Extract the all tagged terms for Protein, SCL, Tissue.
end
Generate a list of all possible triplet combinations **D**.
for *each triplet combination in* ***D*** **do**
  for *each abstract* **do**
    Count co-occurrence $\delta_{ak}(T,P,L)$.
  end
  for *each sentence* **do**
    Count co-occurrence $\delta_{sk}(T,P,L)$.
  end
end
for *each triplet combination in T* **do**
  Calculate the weighted co-occurrence score, Equation (6.5).
end
Convert Co-occurrence score to Z-scores, Equation (6.6).

6.1.6 Experimental design and evaluation

To access the performance of our text mining system on the extraction of the protein tissue-specific SCL, we conducted two types of analysis.

- to assess the validity of our approach, we performed an intrinsic evaluation using a gold- standard benchmark dataset of protein tissue-specific SCL of ten tissues.
- to assess the utility of our text mining tool, we compared the results of our approach with entries in a popular experimental database for fourteen cell lines.

Mappings of tissue terms and SCL terms

In the literature, the information of tissues, cell lines, and protein SCLs are mentioned and unified in different details which depend on the interest of researcher and the power of experimental detections. For example, the SCL of protein C-terminal binding protein 1

(CTBP1) has been detected locating in Nucleoplasm whereas the SCL was mentioned in more generic term Nucleus in the literature [178]. To efficiently assess the performance of our text mining information extraction system, we imported the annotations from BTO, Entrez Gene[179] and GO cellular component, for tissue, gene/protein, and SCL annotations respectively. Due to the hierarchical nature of the ontology, it is necessary to select a subset of terms to be used as the basis for the assessment. In case one term was a child term of another, we selected the broader parent term through *is_a* and *part_of* relationships, see Figure 6.3. In practice, the relations among SCLs accessed by the parental relation of gene ontology. These relations are represented in a graph structure. We use the QuickGO REST application programming interface (API) which provides access to the ancestors of a GO term of interest from QuickGO.

```
https://www.ebi.ac.uk/QuickGO/services/ontology/go/terms/[query
GO identifier] /ancestors?relations=is_a%2Cpart_of
```

In the end, ten representative SCL terms were chosen to use as the final assessment and later for evaluation.

A similar process was performed to the various types of tissue and summarized ten tissues for the evaluation. To tackle on the relations of tissues and cell lines via BTO, we used an in-house script to query the relation from the BTO XML file at `https://bioportal.bioontology.org/ontologies/BTO`.

Establishing gold standards

Following the above text preprocessing, a reference set based on the manually curated MEDLINE abstracts were collected. The abstracts are lacking information about the relationship between tissue, gene and SCL were excluded from the benchmarking set, as well as the abstracts with incomplete information about one of them were excluded. The reference set comprises 170 diverse biomedical publications which contain tissue, gene, and SCL. The curation produced a positive reference set of 220 associations between 132 genes and ten tissues and ten SCLs. We defined the negative set as all other 12980 possible triplets of the same genes, tissue and SCL in the benchmark.

Integration of experimental data resource

So far, there is not yet a tissue-specific SCL dataset in public databases. Hence, we are not able to perform an extinct evaluation of the text-mined results of protein tissue-specific SCL. However, HPA provides the cell line information in which the SCLs of protein were

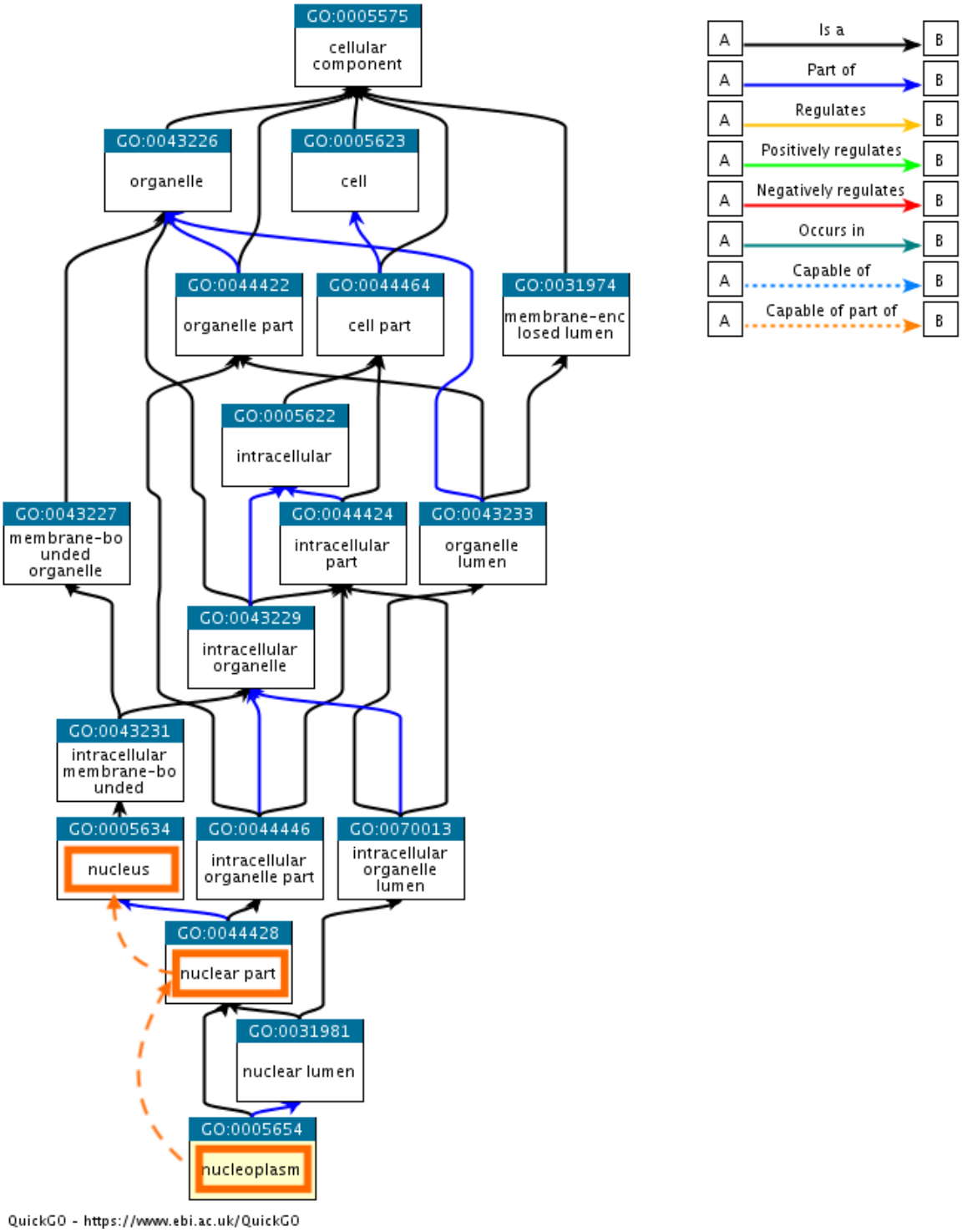


Fig. 6.3 SCL mapping along the GO tree. An example of the mapping from the specific GO-CCO term Nucleoplasm to the generic but representative GO-CCO term Nucleus along the GO Ancestor Chart.

detected in the cellular atlas [6]. To demonstrate the utility of our approach, we perform the information extraction of protein cell line-specific SCL. For that purpose, we generated a dictionary of human cell lines using the same approach as for the dictionary of human tissue. Thirty SCL terms listed in HPA was summarized to the ten SCLs and their GO identifiers Section 6.1.6. Whereas, we take directly the fourteen specific cell lines which are listed in HPA for the final evaluation.

Performance metric

The text-mined triple associations are associated with a normalized score Equation (6.6). We next ranked these association by descending scores and compared them to the reference set. The results were presented in receiver operating characteristic (ROC) curves by plotting the true positive (TP) rate as the function of false positive (FP).

Besides, we also evaluated our text mining system using the following standard information retrieval metrics: recall, precision, and F measure (F1 score) defined in below:

$$\begin{aligned} Recall &= \frac{TP}{TP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ F1 &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \end{aligned} \tag{6.9}$$

where TP, FP and FN stand for the number of true positives, false positives and false negatives, respectively. The true positives and false positives in the text mining results are the correct and wrong triplets based on the text corpus, respectively. The false negatives are the true triplets occur in the text corpus which do not occur in the text mining results.

6.2 Results

6.2.1 Dictionary-based tagger

Since the number of abstracts in MEDLINE is quickly growing every year, an efficient tagging component for genes, tissues, and SCLs are necessary. The tagging component in this work is implemented in Java 8 and accessible via API. Given a PubMed identifier PMID, it returns annotations in the JavaScript Object Notation (JSON) format.

The tagging component combines PubTator which is the state-of-the-art tool for tagging gene mentions (see Section 6.1.3) and the dictionary-based method for identifying the tissueSCL in the text. Since the dictionaries already contain all relevant synonyms and

morphologic variations of terms from the dictionary construction, it is sufficient to exactly match parts of the text to existing dictionary entries. For this purpose, all dictionary entries were tokenized and organized into trie structure (prefix tree) for quick lookups. The well-established NLP tool, the Stanford CoreNLP toolkit [180], was used to perform the tokenizations. The same tokenization process was also applied to the retrieved abstracts from MEDLINE. For each textual token in the text, a lookup in the trie is performed to search for the possible matches. If multiple matches are found in the same location, the tagger selects the longest subsequence to avoid ambiguous tagging results (e.g. "skin fibroblast" would be tagged as "skin fibroblast" [BTO:0001255], instead of "skin" [BTO:0001253]).

The tagger is then applied to automatically annotate a large number of MEDLINE abstracts, generating annotations on tissues, genes, and SCLs. The annotated corpus is formatted in JSON for further analysis.

6.2.2 Evaluation against manual curated corpus - Tissue

The performance of the proposed approach is first evaluated for tissue-protein-SCL triplet extraction from the biomedical literature on the manually curated benchmark datasets, described in Section 6.1.6. This dataset consists of a manually annotated list of tissue-protein-SCL triplets from 170 MEDLINE abstracts which contain 220 ground truth triplets for ten tissues and ten summarized SCLs. Using these benchmark datasets, we report the accuracy of our approach with standard measures (precision, recall and F-measure Equation (6.9)).

Here we focus on benchmarking effort on accessing the quality of the result of extracting tissue-protein-SCL triple associations from given literature. We, therefore, compare the text-mined triplets to the manually curated corpus, considering all detected and top scored triple associations.

Figure 6.4 show that our text mining system can extract tissue-gene-SCL associations with high specificity (low false positive rate) and sensitivity, 0.88 and 0.84 respectively. The optimal threshold to define the positive triplet is -0.034 according to our benchmark dataset. Since our co-occurrence approach utilizes different entity tagging tools, we obtain rather high recall 0.724, fair precision 0.65, and F1 score 0.724, see Table 6.1. From the distribution of the normalized z-scores of all text-mined triplets, shown in Figure 6.5, with the higher threshold score, the high precision (more true positive triplets) can be achieved.

Table 6.1 Performance of text mining system for triplet prediction

Optimal threshold	Accuracy	Precision	Recall	F1 score	ROC-AUC
-0.034	0.830	0.650	0.818	0.724	0.820

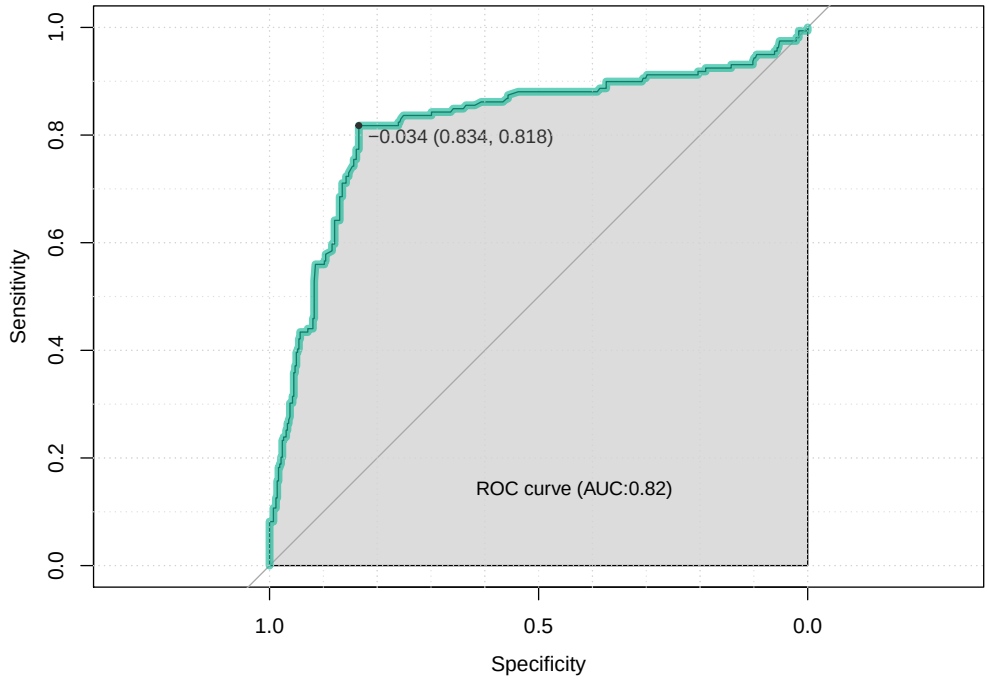


Fig. 6.4 Benchmark of tissue-protein-SCL association obtained through text mining.

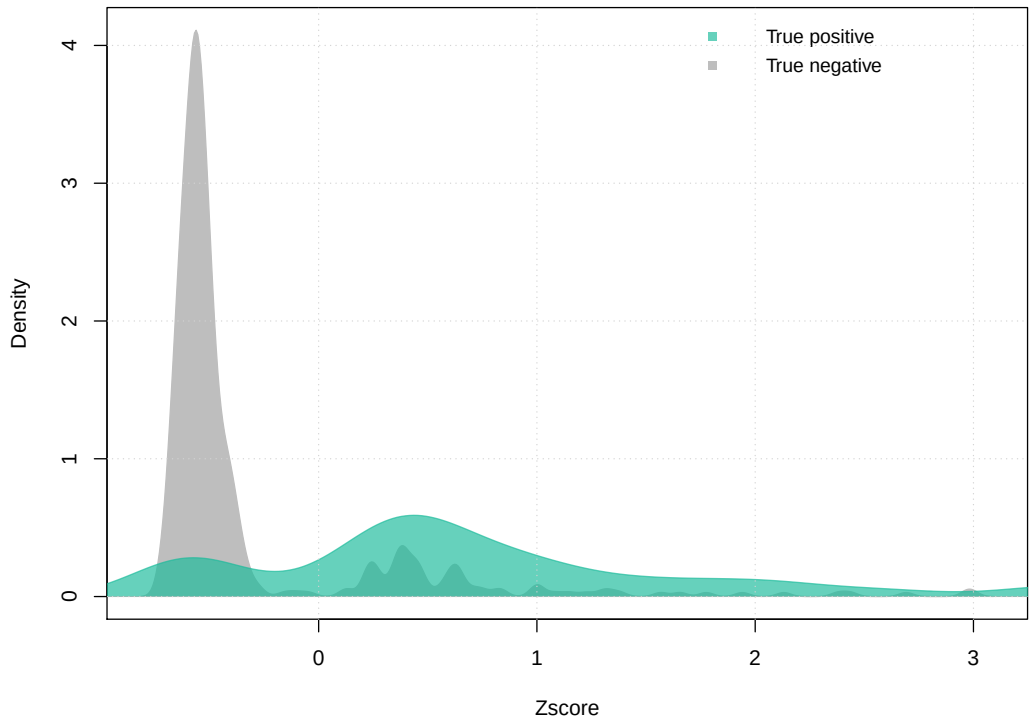


Fig. 6.5 Histogram bar chart of scored true positive triplets and negative positives.

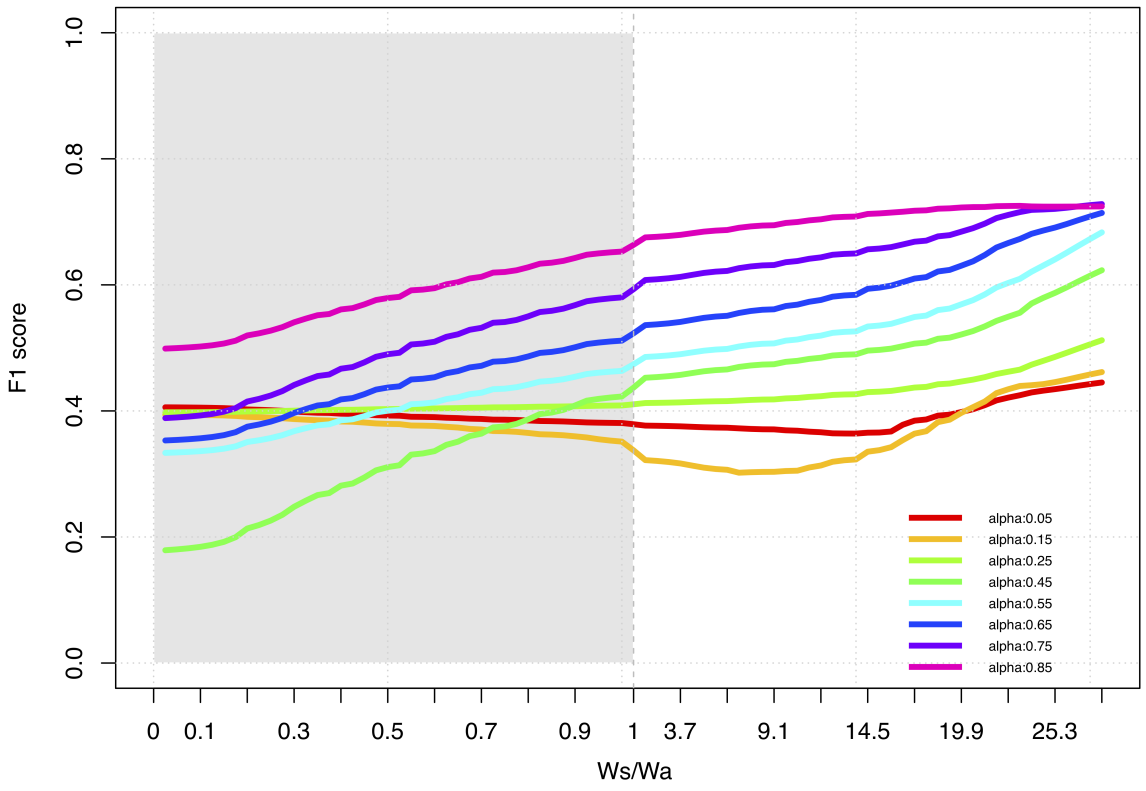


Fig. 6.6 Tuning the scoring parameters.

6.2.3 Evaluation against experimental dataset - Cell lines

To assess the potential of our approach in assisting database curation, we performed an extrinsic analysis by comparing our text-mined results against experimentally curated relationships for a total of fourteen cell lines. HPA, the human protein atlas, is a database of human annotation data produced by wet-lab experiment. Its scope includes the pathology of human cancer transcriptome, tissue-restricted expression of proteome and transcriptome in the major tissues of the human body, and the spatial distribution of proteome at the subcellular level detected by the immunohistochemical staining using antibody [28]. Data collection from HPA is explained in detail in Section 6.1.

The mined results were compared with the cell line specific experimental data from HPA. As shown in Figure 6.7, the red bars denote the counts of text-mined results for each cell line, and the blue bars denote the counts of curated variants for each tissue in the HPA dataset. From HPA, we collected 44000 individual cell line-protein-SCL triplets for seventeen cell lines. We extracted 32350 cell line-protein-SCL from the literature using our text mining approach 6736 human proteins locating in eleven SCCs. As is apparent in the figure, for most of cell lines, the proposed text mining system extracts a significantly larger number of triplets than the existing triplets in HPA. However, for the cell lines U2-OS, U-251 MG, and A-431, many associations cannot be found by the proposed text-mining approach.

Evaluation of overlapping triplets First, we evaluated the accuracy of the <cell line, protein, SCL> triplets which can be found in both HPA database and our text-mining system. 902 triplets are overlapped from two data sources. In this category, a correct association has already been confirmed by the presence of this association in HPA. Thus, the remaining step to confirm the correctness of the full triplet is to assess whether our threshold (-0.034) (see Figure 6.8) can efficiently distinguish the positive associations from the background noises. In the end, we were able to match triplets with an average accuracy of 0.83 for 667 proteins in ten SCLs expressed over fourteen cell lines, see Table 6.2.

The un-curated triplets HPA has reported 10658 human proteins and their subcellular distribution in different cell lines. Our text mining approach returned 31468 triplets that were not found in HPA. It includes 20158 new triplets for 4043 proteins occurring in HPA and 11732 new triplets for 2812 human proteins which have no experimental detection data in HPA. These associations and their supporting literature references are potential candidates for curation. On the other hand, it is possible that these results also contain false positives. Thus further manual curation is required.

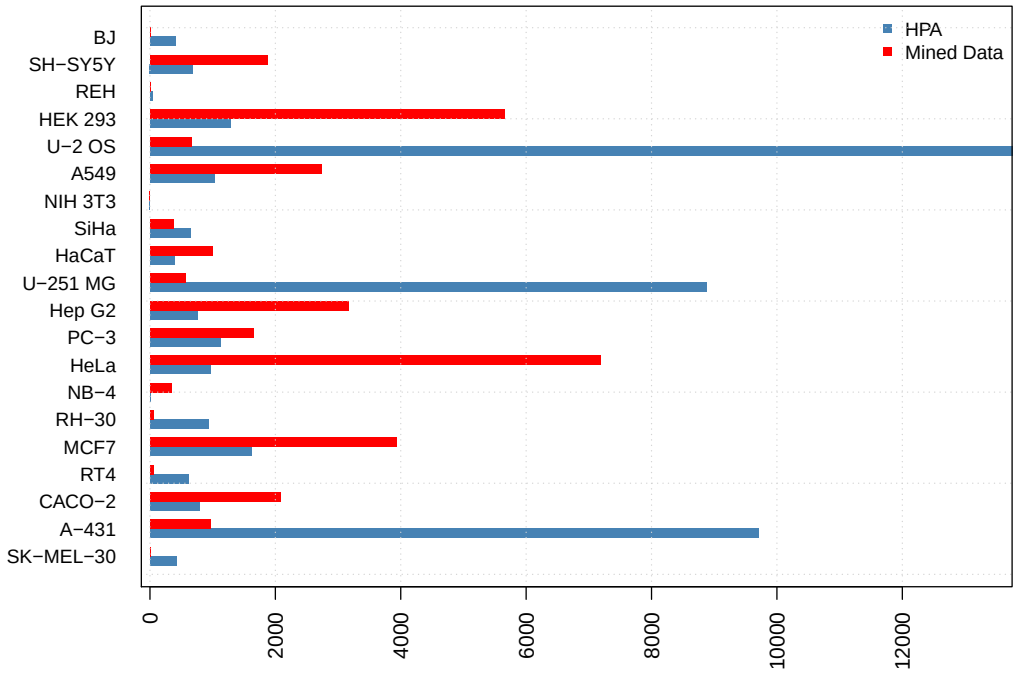


Fig. 6.7 Comparison of the text-mined results with HPA experimentally validated cell line data.

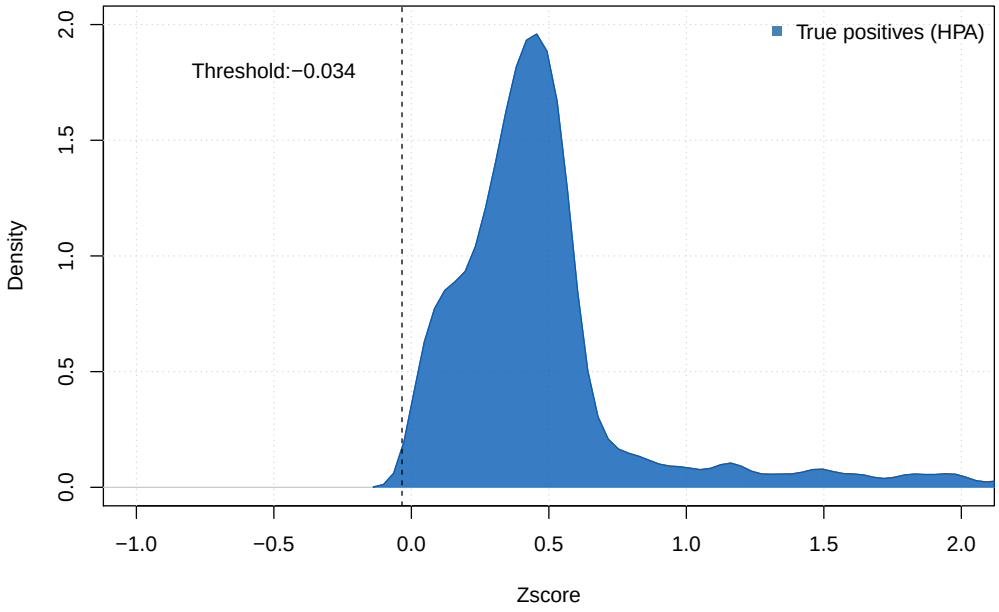


Fig. 6.8 Distribution of the scored triple association.

Table 6.2 Accuracy of overlapped triplets.

Cell lines (BTO identifier)	Accuracy
A-431 (BTO:0000017)	0.842
U-2 OS (BTO:0001938	0.915
U-251 MG (BTO:0002035	0.880
MCF7 (BTO:0000093)	0.738
CACO-2 (BTO:0000195)	0.871
HEK 293 (BTO:0000007)	0.667
RH-30 (BTO:0005379)	1.000
Hep G2 (BTO:0000599)	0.873
PC-3 (BTO:0001061)	0.833
SH-SY5Y (BTO:0000793)	0.950
A549 (BTO:0000018)	0.780
HeLa (BTO:0000567)	0.863
HaCaT (BTO:0000552)	0.875
SiHa (BTO:0002210)	1.000
Average	0.863

The triplets present only in HPA However, 41256 triplets detected in HPA were not returned by our text mining approach. The reason why these triplets were not returned by our text mining approach is likely that they are not mentioned in the abstract, and more likely are mentioned in 'Material' section in the full text. We assume that most of the experimentally validated triplet published in HPA was not mentioned in literature, especially the cell lines. Although the proteins could be specifically expressed in a certain number of tissues and cell lines, the HPA mainly use U2OS cell line for the convenience of experimental manipulation to access the possible SCL of protein in the cell.

6.2.4 Creation of TS-SCL database

A large-scale text mining was performed against 127,7785 abstracts. A database based on the top scored predicted results, and integrated tissue-protein-SCL associations from HPA experimental resource was therefore established. Although we mostly emphasis on the text mining aspects in this chapter, the TS-SCL database integrates <Tissue, Protein, SCL> associations from several data sources. The data sources include the triple associations from 1. the large scale text mining for human tissues and cell lines (Text mining channel); 2. the manually curated data (Knowledge channel); 3. the experimental data from HPA database (Experiment channel); 4. the prediction results using tissue-specific-BCMRFs

Table 6.3 Overview of TS-SCL database.

Data resource	Reliability	Cell line & Tissue	Proteins	SCLs	Associations
Experiment channel					
Human Protein Atlas	Approved	19	5626	11	21461
	Supported	19	3492	11	13914
	Validated	17	1540	11	6763
	Total	19	10658	11	42138
Text mining channel					
Tissue	> 3	10	538	11	982
	$1 \sim 3$	10	1971	11	3591
	$0.5 \sim 1$	10	3591	11	5834
	$0 \sim 0.5$	10	4019	11	11608
	$< 0^{*}$	10	11222	11	83688
	Total	10	5925	11	22015(105703**)
Cell line	> 3	18	303	11	486
	$1 \sim 3$	17	1841	11	2942
	$0.5 \sim 1$	17	1007	11	1686
	$0 \sim 0.5$	19	440	11	725
	$< 0^{*}$	18	2844	12	28044
	Total	19	2521	11	5838 (33883**)
Knowledge channel		10	132	10	220
Prediction channel		9	243	8	1643
Total					71854(183587**)

* Low confident associations which are not included in the database.
** Number of total mined associations.

algorithm (Prediction channel). The integration of heterogeneous data sources overcomes the shortcomings of each resource.

Table 6.3 provides an overview of the total evidence landscape of the database, showing that the both experimental data and text mining pipeline are the largest contributors of associations. However, it is important to note that this number depends strongly on the confidence cutoff of Z score. We use a modest cutoff value of 0 to ensure the precision. Table Table 6.3 lists some of the key characteristics of the extractions from the 127, 7785 PMIDs. It is noteworthy that triplets were extracted from only 19190 of the abstracts. This low coverage is often due to that the 'abstract' hints to a possible tissue and SCL association, but the specific details are contained in the full-length article. 73436 triplets were found across all abstracts, among which 31072 was unique. All associations from all evidence sources are available for download in tab-delimited format at `http://agbi.techfak.uni-bielefeld.de/tsscldb/`.

6.2.5 TS-SCL database web interface

Although the tab-delimited text files are convenient enough to manipulate for bioinformaticians for large-scale analyses, a web interface is much more intuitive for researchers interested in individual gene/protein, tissue or SCL. For this purpose, we developed a user-friendly interface for the TS-SCL database resource to support various types of searches. It allows users to either query for a gene/protein identifier to find associated tissue and SCL or query for a tissue to find associated genes, see Figure 6.9a. In either case, the results are displayed in two different ways: the triplet-view and the PMID-view. The triplet-view (see Figure 6.9b) shows the extracted triplets and additional information for these triplets. A table of triplet associations indicating data resource (Knowledge, Experiments, Text mining and Predictions) is presented to users. Besides summarizing the imported information, it provides direct hyperlinks to the source entries in the external databases. The PMID-view (see Figure 6.9c) shows the information at an abstract level with detailed annotation of matching tissue, protein/gene, and SCL.

Currently, the stored results and the web interface can be found at: `https://agbi.techfak.uni-bielefeld.de/tsscldb/`. We have conducted the following query into our database to demonstrate the usability of the system. Search the database for the protein via its gene name TP53. A search for TP53 retrieved all available <Tissue,Protein,SCL> triplets across tissues in TS DB. Figure 6.9b shows a screen shot of the displayed results. The tissue, protein, SCL extraction zone and the estimated Z scores are shown in a tabular format. The entire result table can be downloaded as a spreadsheet with the Download results link. The last column shows the number of the publications which support this triple associations. By

clicking on the button, a new window tab is opened up, and shows the list of abstracts. The actual abstract text with tagged keywords for tissue, protein, and SCL can be shown and be hidden by clicking on the title, see Figure 6.9c.

The web services was implemented in Python 3 [167] using Django web framework [181] integrating SQLite [182].

6.2.6 Generality of the approach

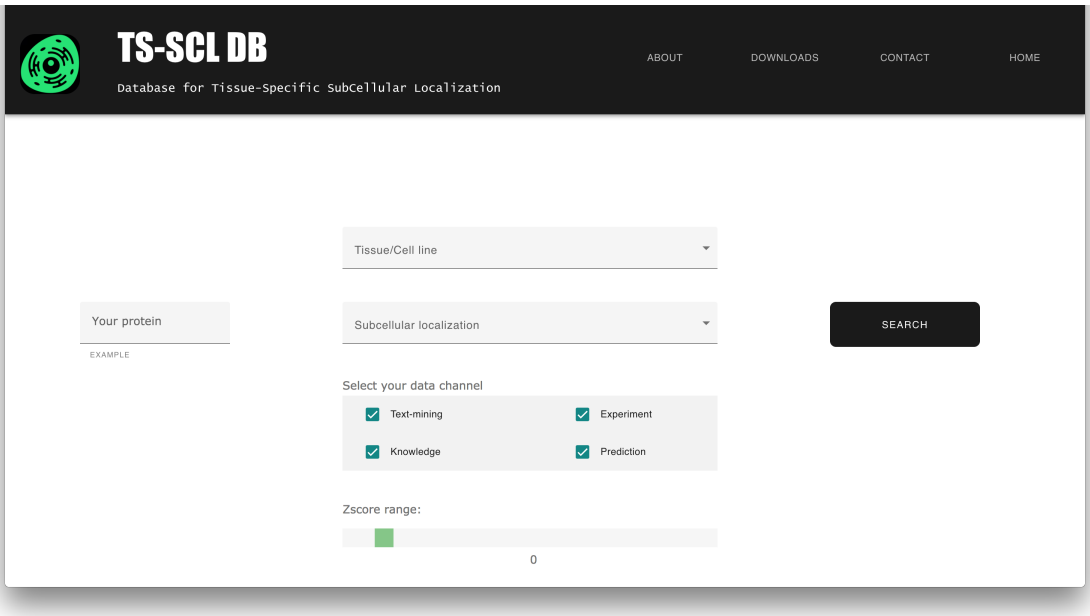
The approach of text mining described in this paper is readily applicable to recognize the named entities of tissue and SCL in different detailed level to fulfill researcher's interest by adjusting the dictionaries and defining the mapping files. For instance, the information extraction of SCL can be general, such as 'Nucleus', and be specific to the sub-unit, such as 'Nuclear membrane', 'Nucleoli', 'Nucleoli fibrillar center'. Similarly, the information extraction of tissue can be general, such as the connective tissue, and be as detailed as 'bone', 'fibroblast', 'adipose' tissue, etc. Combining the tagger with the co-occurrence scoring scheme for information extraction (IE) is equally flexible. As previously mentioned, the scoring function was initially developed to extract functional associations between paired proteins for use in the STRING database based on co-occurrence of gene names within biomedical literature [160]. The scoring schema has been successfully applied in disease–gene associations and so on [183, 148, 177]. Together with our results, we are very confident that our adapted scoring schema for triple association can be applied to other biological contexts, such as disease-specific SCL.

6.2.7 Limitation and future direction

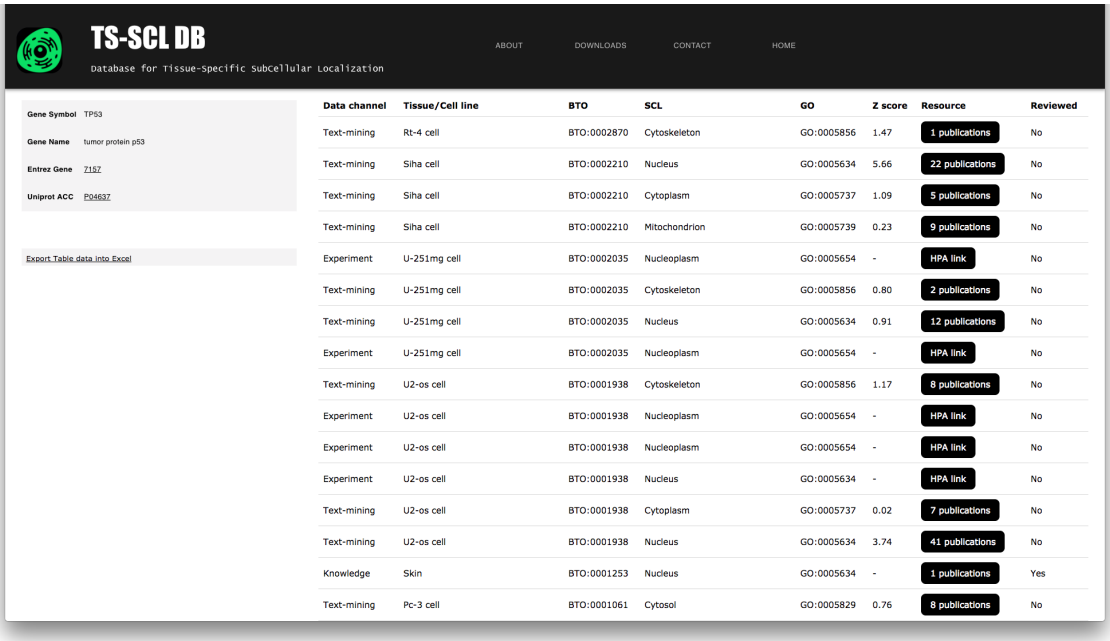
In this section, we discuss the limitations of our approach, identify areas of work that may enhance our approach and improve its utility for future research and other applications.

One drawback of co-occurrence methods is that they are unable to extract the direction of an association and have difficulty distinguishing between direct and indirect associations, and the negative association. Another significant limitation of the current approach is that it mines information only from abstracts and not full text or supplementary material, which have been shown to be an important source of tissue/cell line and SCL information. An extension to full text will require more advanced systems to overcome the additional noise in the full text and tables.

During the process of literature manual curation, and evaluation, we observe that the research of protein SCL is often carried out with a focus on a particular disease. Moreover, a set of cell lines are suitable to use for studying one disease, and some cell lines are used

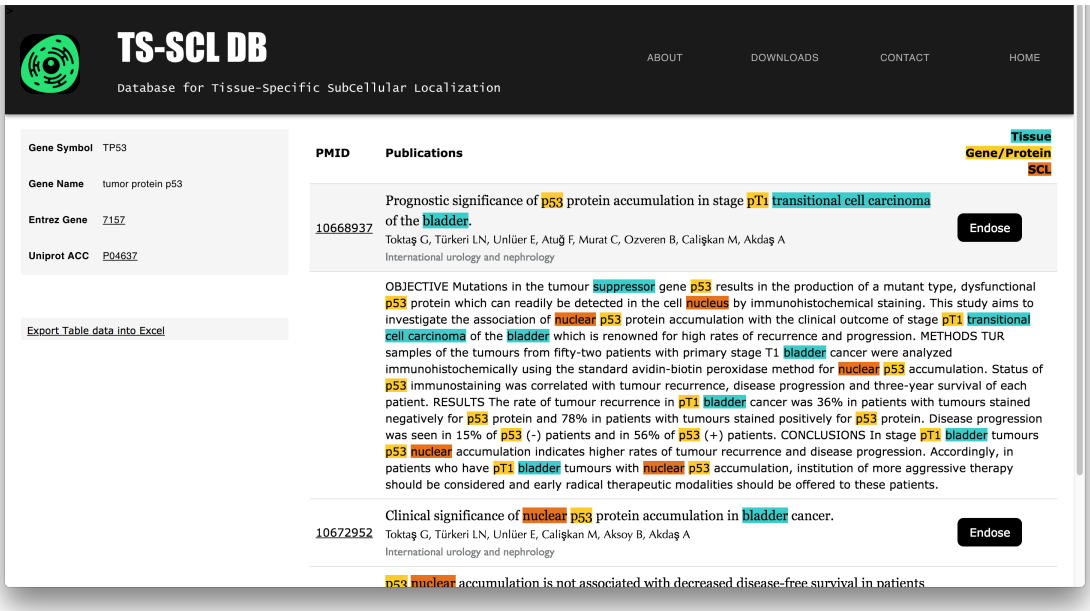


(a) Three search fields.



Data channel	Tissue/Cell line	BTO	SCL	GO	Z score	Resource	Reviewed
Text-mining	Rt-4 cell	BTO:0002870	Cytoskeleton	GO:0005856	1.47	1 publications	No
Text-mining	Siha cell	BTO:0002210	Nucleus	GO:0005634	5.66	22 publications	No
Text-mining	Siha cell	BTO:0002210	Cytoplasm	GO:0005737	1.09	5 publications	No
Text-mining	Siha cell	BTO:0002210	Mitochondrion	GO:0005739	0.23	9 publications	No
Experiment	U-251mg cell	BTO:0002035	Nucleoplasm	GO:0005654	-	HPA link	No
Text-mining	U-251mg cell	BTO:0002035	Cytoskeleton	GO:0005856	0.80	2 publications	No
Text-mining	U-251mg cell	BTO:0002035	Nucleus	GO:0005634	0.91	12 publications	No
Experiment	U-251mg cell	BTO:0002035	Nucleoplasm	GO:0005654	-	HPA link	No
Text-mining	U2-os cell	BTO:0001938	Cytoskeleton	GO:0005856	1.17	8 publications	No
Experiment	U2-os cell	BTO:0001938	Nucleoplasm	GO:0005654	-	HPA link	No
Experiment	U2-os cell	BTO:0001938	Nucleoplasm	GO:0005654	-	HPA link	No
Experiment	U2-os cell	BTO:0001938	Nucleus	GO:0005634	-	HPA link	No
Text-mining	U2-os cell	BTO:0001938	Cytoplasm	GO:0005737	0.02	7 publications	No
Text-mining	U2-os cell	BTO:0001938	Nucleus	GO:0005634	3.74	41 publications	No
Knowledge	Skin	BTO:0001253	Nucleus	GO:0005634	-	1 publications	Yes
Text-mining	Pc-3 cell	BTO:0001061	Cytosol	GO:0005829	0.76	8 publications	No

(b) View of available triple associations.



(c) View of the annotated document.

Fig. 6.9 Illustration of web interface.

in more general manner. A defective translocation of protein (mislocalization) is one of the reasons which causes disease. We have an insight into the necessity of the data curation of protein disease-specific SCL. Our flexible dictionary-based text mining system can be used to curate such data by integrating a dictionary of disease and adjusting dictionary of tissue.

6.3 Summary

In conclusion, we have shown that our dictionary-based approach for extraction of <tissue, protein, SCL> from MEDLINE abstracts is successful. The intrinsic evaluation shows that our approach achieves good precision and recall. Our comparative analysis of the experimental data confirms the accuracy of our approach and demonstrates that text-mined results may be potentially useful for expanding the coverage of curation and improving curation quality.

The proposed text mining system was validated by applying to extract tissue-specific SCL associations from a broad set of abstracts in MEDLINE. The extracted tissue-specific SCL associations are stored together with the abstracts from which they were extracted, as well as with additional relevant information that was obtained from these abstracts. A web-based

database for the obtained data was established. Finally, we would like to extend our system to run on full-length articles in the further.

Chapter 7

Tissue-specific subcellular distribution of the human AGO2 protein

In the previous chapters, we introduced BCMRFs algorithm to address the general multi-SCL problem. BCMRFs integrates the factors of the physical interactions among proteins, the physical spacial adjacency of SCCs and the protein sequence , see Chapter 4. In Chapter 5, we discussed the importance of understanding the tissue-specific SCLs of protein. We applied the BCMRFs on tissue-specific PPINs to predict tissue-specific SCL of human proteins. Furthermore, we developed a scoring model based text mining system and extracted tissue-specific SCL associations from the abstracts of a large number of biomedical papers, see Chapter 6.

For each task mentioned-above, we have already evaluated and validated with benchmark sets. Also, we performed large-scale prediction on human proteome and generated a significant amount of results. However, for the researcher who often focuses on the small number of proteins, a small-scale analysis would be more precise and accessible. In this chapter, we demonstrate how to use our methods to analyze the tissue-specific subcellular distribution of a protein of interest. The computationally predicted tissue-specific SCL results can be helpful to generate assumptions about the novel function of the protein, and to understand the cellular mechanisms. Besides, the TS-SCL DB which contains text-mined data can be used to validate the previous prediction, search for other possible tissue-specific SCLs, and retrieve the relevant literature supports.

The members of Argonaute (AGO) protein family are the key players in gene-silencing pathways guided by small noncoding RNAs, including short interfering RNAs (siRNAs), micro RNAs (miRNAs) and Piwi-interacting RNAs (piRNAs). AGO proteins are universally expressed in many organisms. Human AGO1, AGO3, and AGO4 genes are clustered on chromosome 1, whereas the AGO2 gene is localized on chromosome 8 [22]. It has been

reported that human AGO2 protein is an endonuclease and a major component of the RNA-induced silencing complex (RISC). Ago2 binds miRNAs, siRNAs, and Piwi-interacting RNAs and mediates the loading of these small RNAs onto RISC to recognize specific targets through base-pairing, ultimately leading to mRNA translation inhibition or degradation [17].

AGO2 has been mostly known as a cytoplasmic protein due to its ectopic expression in the cytoplasm and is distributed in the cytoplasmic RNA granules, including P-bodies and stress granules [22]. However, later studies and data suggested that AGO2 might be a nuclear-cytoplasmic shuttling protein in cells and may be involved in various nuclear processes [25]. Recent evidence demonstrates that the nuclear distribution of AGO2 occurs in a cell type- and tissue context-dependent manner and may correlate with its various functions in the regulation of gene expression [17].

To date, the subcellular distribution of the human AGO2 protein, its tissue specificity and its function remains as essential subjects of scientific debate. In next sections show how to use the tissue-specific SCL prediction workflow (see Figure 5.1) to analyze the subcellular distribution of the human AGO2 proteins in various cell lines and tissues.

7.1 Tissue-specific PPI networks of the human AGO2 protein

Following the workflow in Figure 5.1, seven tissue-specific PPINs on which the human AGO2 protein shows significant expression are generated. The seven human tissues include skin, brain, uterine cervix, liver, colon, lung and kidney. The tissue-specific networks can be used to unravel the distinct functions of MFP by examining the neighbors of the protein of interest [10]. Thus, seven tissue-specific PPINs were analyzed and compared to reveal the tissue-specific function of AGO2 across the seven tissues. The Figure 7.1 shows that the interacting neighbors of human AGO2 protein are notably different in the seven tissues. The various interacting partners across human tissues suggests that AGO2 might carry multiple functions in different tissues.

The seven tissue-specific (tissue-specific) PPINs are restricted to the physical protein interactions which can be used to infer the SCLs of protein from the SCLs of the interacting neighborhood. Thus, the clear difference between the interacting partners across seven tissue can imply that AGO2 might be a tissue-specific MLP. In Figure 7.2, we observed the significant difference of the subcellular distributions of all SCLs of the neighborhood of AGO2 protein across tissues. Among thirteen different SCLs in the interacting neighborhood, the nuclear distribution, and cytoplasmic distribution take a large proportion in the most of

Table 7.1 Interacting partners of AGO2

Gene symbol	Function	Co-expressed tissues
DICER1	Small RNA generation and RISC loading [184]	skin, lung, liver, colon, kidney, cervix, brain
MTDH	Facilitation RISC activity [185]	cervix
PRKRA	Regulate Dicer-mediated dsRNA processing [186]	colon, kidney
TNRC6A TNRC6B	GW protein family. Coordination of downstream silencing events [186]	liver, kidney colon
UBR5	Help in the recruitment of downstream factors by binding to GW proteins [187]	liver, kidney, lung
UPF1	Support AGO complexes in target association [187]	skin, liver, kidney, cervix
IPO8	Stabilization of AGO–mRNA interactions [188]	lung, cervix

the tissues. The difference of the SCL distributions indicates again that the potential multi-localization and the distribution are tissue dependent. Furthermore, the multi-localization of AGO2 also implies its multi-functional role.

7.2 Characterization of the tissue-specific networks

Proteins often interact with other molecules to carry out their specific functions under various context. The interactions might induce the potential translocation of protein or protein complex. In this section, we discuss the relationship between the human AGO2 protein and its interacting partners, and the relevance of their biological functions.

Proteins in the network neighborhood that directly interact with AGO2 are relevant for regulating RNA silencing event, and can also include proteins essential for regulating RNA silencing machinery or other functions. Table 7.1 shows some examples of the direct interacting partners of AGO2 functions, and the co-expressed tissues. Below, we categorize the essential interacting partners of AGO2 regarding their functions together with AGO2.

7.2.1 Roles in RNA silencing event

endoribonuclease dicer protein (DICER1) is an RNase III family endonuclease that processes double-stranded RNA and precursor miRNAs into siRNAs and miRNAs, respectively. DICER1 is universally expressed in all seven tissues and interacts directly with AGO2

through the Piwi domain of AGO2. DICER1, AGO2, heat shock protein 90 (HS90) and RISC-loading complex subunit TARBP2 (TARBP2) constitute the trimeric RISC loading complex (RLC) [186, 189].

HS90 is another interacting partner of AGO2. The general function of HS90 is a molecular chaperon that promotes the maturation, structural maintenance and proper regulation of specific target proteins involved for instance in cell cycle control and signal transduction. The stable binding between AGO2 and DICER1 was dependent on the activity of HS90 protein by binding AGO2 [190, 191].

interferon-inducible double-stranded RNA-dependent protein kinase activator A (PRKRA), is also known as PACT. In mammals, PRKRA interact with DICER1 and AGO2, and have been implicated in miRNA biogenesis [192].

LYRIC protein (MTDH), also known as AEG-1, is required for optimum RISC activity facilitating small interfering RNA and micro RNA-mediated silencing of luciferase reporter gene. Previous coimmunoprecipitation and colocalization studies confirmed that MTDH is also a component of RISC [185].

GW proteins (including TNRC6A, TNRC6B and TNRC6C) are characterized by an N domain containing multiple glycine–tryptophan (GW) repeats. This domain directly interacts with AGO proteins and is therefore referred to as the AGO-binding domain. GW proteins play a role in RNA-mediated gene silencing by both miRNAs and siRNAs [186, 184].

importin 8 (IPO8) is a gene silencing factor that targets AGO2 to distinct mRNAs and stabilize AGO–mRNA interactions [188]. Likewise, protein regulator of nonsense transcripts 1 (UPF1) participates in RNA silencing by facilitating the binding of the RISC to the target and by accelerating the decay of the mRNA [187]. Furthermore, IPO8 mediates the import of human AGO2 along with associated microRNAs into the nucleus via [193].

Most of the interacting partners of AGO2 mentioned-above play roles in RNA silencing event. However, some interactions universally happen across seven tissues, such as the interaction with DICER1. The other interactions are, by contrast, tissue-dependent, such as the interaction with MTDH which only takes place in uterine cervix tissue, whereas PRKRA in colon and kidney. Moreover, Eystathioy [194] revealed that TNRC6A (GW protein) is ubiquitously expressed in different tissues including heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. However, the tissue-specific networks suggest that the interaction between AGO2 and TNRC6A might occur only in liver and kidney tissue but not brain tissue. Another explanation would be that such interaction has not been experimentally proven yet in the other tissues.

7.2.2 Roles in mRNA splice and translation

probable ATP-dependent RNA helicase DDX20 (DDX20) and Gem-associated protein 4 (GEMI4) are members of the the survival of motor neurons complex (SMN complex) which plays a catalyst role in the assembly of small nuclear ribonucleo proteinss (snRNPs), the building blocks of the spliceosome. Therefore, it plays a crucial role in the splicing of cellular pre-mRNAs. Moreover, AGO2 also interacts with pre-mRNA processing factor pre-mRNA-splicing factor ATP-dependent RNA helicase DHX15 (DHX15) which is involved in disassembly of spliceosomes after the release of mature mRNA. The interactions between AGO2 and SMN complex proteins and DHX15 indicate that AGO2 plays a role in the regulation of mRNA splicing event. Furthermore, both AGO2 and interactor eIF-2-alpha kinase activator GCN1 (GCN1) has demonstrated their roles in the regulation and activation of mRNA translation [26, 195].

7.2.3 Roles in tumorigenesis

The p53-related protein p63 has pleiotropic functions, including cell proliferation, survival, apoptosis, differentiation, senescence, and aging [196]. The involvement in the tumor-suppressive mechanism of apoptosis made p63 be recognized as one of the most important tumor suppressor protein. Since miRNAs/oncogenes/tumor suppressive genes are critical in tumorigenesis. The association of AGO2 and p63 protein imply that as an essential mediator of miRNAs function and maturation, AGO2 can undoubtedly affect tumorigenesis.

7.3 Analysis of the prediction results

The seven tissue-specific PPINs in Section 7.1 were used to perform tissue-specific SCL prediction as shown in Figure 5.1 for AGO2 protein. For each tissue, we obtain the a set of confident SCLs propositions. Following that, we compare the tissue-specific SCL prediction results with the generic SCL annotations of AGO2. In addition, we use our text mining system to validate, and search for more possible tissue-specific SCLs of human AGO2 protein which are not captured by our previous prediction.

7.3.1 Generic SCLs

Both the knowledge-based and the experimentally detected SCL annotations of human AGO2 protein from various data resources can be retrieved via UniProt API interface. Table 7.2 shows the SCL annotations of AGO2. However, during the prediction and evaluation, we

exclude the particular annotations such as 'complex', and only consider the subcellular compartment related annotation. The major SCLs of AGO2 are therefore nucleus, nucleoplasm, cytoplasm, cytosol, cytoplasmic P-body, polysome, cell junction and extracellular exosome. We summarize the nucleoplasm which is sub-unit of the nucleus as nucleus for the prediction. Likewise, cytosol and extracellular exosome are summarized as cytoplasm and extracellular space respectively. Polysome is a multi ribosomal structure representing a linear array of ribosomes held together by messenger RNA. The summary results in six distinct SCL annotations for AGO2 protein.

Figure 7.3 shows the predicted subcellular distribution of AGO2 across seven tissues. Despite the tissue specificity, our prediction results propose ten SCLs from thirteen possible SCLs from the neighborhood in the networks. Among the summarized SCL annotations from Table 7.2, five are correctly predicted based on the ground truth which mentioned above, which are the nucleus, cytoplasm, P-body and extracellular space and polysome.

7.3.2 Tissue-specific SCLs

From Figure 7.3, we can easily see the distinct subcellular distribution of AGO2 in different tissues. In below, we discuss the roles of AGO2 in different tissues or cell type according to the subcellular distributions. Also, the text mining system can help us to collect the corresponding literature supports and suggest other tissue-specific SCLs which are missed by the prediction. For doing so, we query our database using protein symbol 'AGO2' as input, while the 'Tissue' and 'SCL' option remain unspecified to retrieve all possible SCLs in all possible tissues.

Nucleus and cytoplasm

The nuclear distribution of AGO2 protein is ubiquitously expressed in all seven tissues. Among them, the nuclear distributions in the skin, uterine cervix, and kidney tissue have been experimentally proven using normal tissue cells and the immortalized cell lines HaCaT, HeLa and HEK293 respectively, which are derived from those three tissues respectively [17].

Likewise, we observe the cytoplasmic distribution of AGO2 in all the seven tissues, which agree with the previous studies that AGO2 is primarily expressed in the cytoplasm in HEK293 cells (origin from kidney tissue) [22, 213], H358 cells (origin from lung tissue) [214].

Sharma et al. [17] have experimentally proven that AGO2 is distributed primarily as a nuclear protein in skin, normal cervix, and cervical cancer tissues, whereas the majority of AGO2 is cytoplasmic in HaCaT cells (origin from skin). Moreover, double immunofluores-

cence analysis for MTDH and AGO2 co-localization in nucleus and cytoplasm in QGY-7703 cells which is from liver tissue in Yoo et al. [185]'s study.

Endoplasmic reticulum and polyosome

Detzer et al. [215] provides strong evidence for the view that co-localization of siRNA and Ago2 in the vicinity of the rough ER in ECV-304 cells (bladder) is related to target inhibition. Nevertheless, the density gradient fractionation of cell organelles experiment shows a lack of co-localization in SKRC-35 cells (kidney) in which RNA interference does not occur after the PS-mediated delivery. Dual cell system to identify important steps of intracellular trafficking of siRNA after polysome mediated delivery that is crucial for its biological activity and which seem to be of general importance for the understanding of the intracellular trafficking and release of siRNA in different cell types.

It was elucidated by Barman and Bhattacharyya [216] that AGO2 protein binds on the cytoplasmic side of ER during the miRNA-driven translation repression event in HEK 293 cells (kidney).

Barman and Bhattacharyya [216] found polysome targeting precedes Ago2 and miRNA interaction and repression of target mRNAs on ER using HEK 293 cells (kidney).

P-body

P-body is currently believed to form as a consequence of miRNA-mediated silencing and are sites where repressed mRNAs accumulate and are subject to degradation or storage [217]. James et al. [202] demonstrated the co-localization of the know RISC protein components including AGO2 into P-body structures in U2OS cells. Hubstenberger et al. [218] also observed that mRNAs bound to AGO2 and strongly accumulate in P-bodies.

Argonaute localization to P-bodies was reported to be controlled by two post-translational modifications (phosphorylation and hydroxylation) and the steady-state level of endogenous AGO2 [219] and uncovered involved in mRNAs repression.

Vesicle and Golgi apparus

The endomembrane system of eukaryotic cells allows the spatial and temporal compartmentalization of macromolecule synthesis, sorting, delivery, and degradation. The main organelles of the endomembrane system are the endoplasmic reticulum, the Golgi complex, trans-Golgi network, endosomes, and lysosomes or vacuoles.

Noticeably, our predicting results reveal AGO2 protein's distribution in Golgi apparatus, ER and vesicles (including endosomes and exosomes) in colon tissue. This result suggests

the possible involvement of AGO2 in the endomembrane system and related intracellular transfer. Several previous research supports this assumption. McKenzie et al. [220] revealed that KRAS regulates Ago2's SCL to multivesicular endosomes in colon cancer cell lines.

The distribution of AGO2 in Golgi apparatus is consistent with previous publications of Cikaluk et al. [221], Tahbaz et al. [222]. Furthermore, Stalder et al. [223] summarized that during siRNAs-mediated RNA silencing event, the 'consumed' RISC might dissociate from the mRNP after slicing and recycle back to the rough ER via the Golgi apparatus in HeLa cells.

Cytoskeleton

Moser et al. [224] showed that AGO2 is localized in the centrosome (cytoskeleton) and in the basal body of primary cilia (brain tissue).

Plasma membrane

According to our results, AGO2 is localized in the plasma membrane in skin, liver, colon and kidney tissue which there is no experimental evidence nor literature support available yet. However, Ghosh et al. [225] has proven that AGO2 protein is a member of cell membrane proteins in YTS cells (blood). However, the role of AGO2 in the plasma membrane and the trafficking of AGO2 to the plasma membrane is not yet clear.

Cell junction

Although we have good coverage of predicted tissue-specific SCLs of AGO2 protein and the ones through text mining, there are still experimental detected SCL which are missed by both approaches. The immunofluorescent experiments from HPA have detected the in cell junction distribution of AGO2 in both A-431 cells (skin) and U-251 MG cells (brain), but not in U-2 OS cells (sarcoma). To date, the function of AGO2 localizing in cell junction of skin and brain tissue is not yet clear. However, together with the evidence of AGO2 localizing in Golgi apparatus, endosome, exosome and plasma membrane, we assume that AGO2 might play an essential role of intracellular trafficking and secretion with siRNA in brain and skin tissue, and probably not in sarcoma tissue.

7.4 Summary

The MLPs may have context-specific functions increasing the functionality of the proteome. These multifunctional genes that participate in distinct cellular processes in different SCCs

in a various tissue context. In this chapter, we focused on the human AGO2 protein which is a major component of RNA-induced silencing complex.

We predict the tissue-specific SCLs of AGO2 protein by applying BCMRFs algorithm to the tissue-specific PPINs. The results help us to understand the subcellular distribution of AGO2 protein across seven tissues. We profoundly analyzed the subcellular distribution of AGO2 which occurs in a cell type- and tissue context-dependent manner and the correlation of its various functions in the regulation of gene expression.

Together with the predicting results of tissue-specific SCLs, the evidence from scientific literature and experiments, the functions of AGO2 protein regarding its tissue-specificity can be concluded as in below.

(1) mRNAs degradation. As a core element of RISC complex, AGO2 could directly initiate the degradation of target mRNAs through its catalytic activity in gene silencing processes guided by siRNAs or miRNAs (Universal)

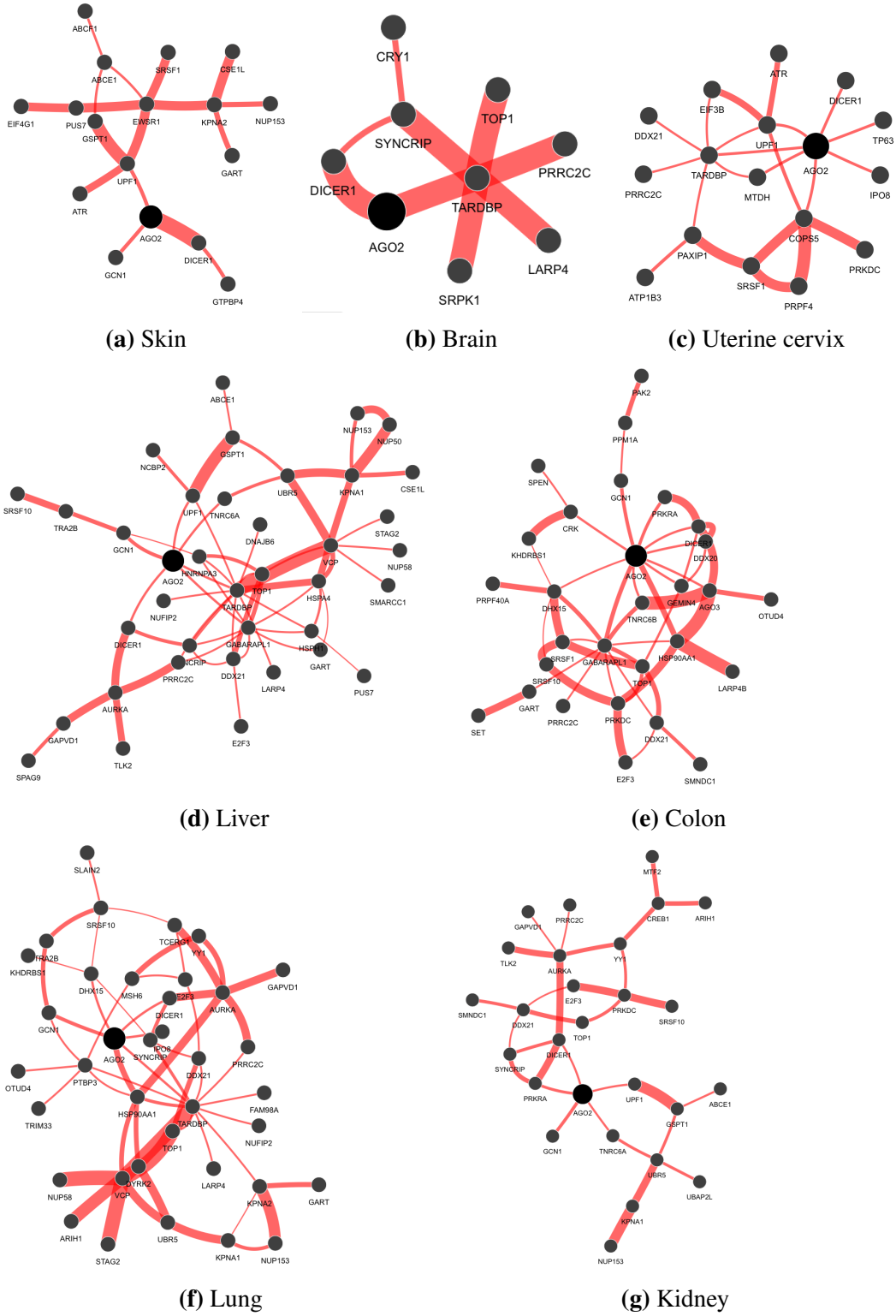
(2) Regulator of miRNAs maturation (DICER universal)

(3) mRNA splicing: Translocation of AGO2 from the nucleus to cytoplasm, polysome (lung and colon, cervix)

(4) Translation repression: AGO2 interacts with GW proteins and localizes in ER, P-body, Golgi apparatus mainly in the uterine cervix, kidney.

(5) Intracellular trafficking and secretion with siRNA: The distribution of AGO2 in the cytoskeleton, vesicle, plasma membrane, cell junction and extracellular space in colon, skin, lung, and liver.

Our text mining system can validate the majority of the prediction results of AGO2 SCLs. However, there are still some prediction results which has no literature support through our text mining system. This might be due to the limitation that the data extraction is only used on abstracts in our text mining system, whereas the tissue, especially the cell line and SCL information might only occur in the main text in the publications. The other explanations would be that the proposed tissue-specific (tissue-specific) SCL is not yet experimentally determine, or that the predicted result is simply a false positive. Either way, these results require further experimental examination.



(a) Skin **(b)** Brain **(c)** Uterine cervix

(d) Liver **(e)** Colon

(f) Lung **(g)** Kidney

Fig. 7.1 Best connected tissue-specific PPINs of human AGO2 protein. The tissue-specific (tissue-specific) PPINs of MLP AGO2 in seven human tissues, including skin, brain,uterine cervix, liver, colon, lung and kidney. The retrieved interacting neighbors of human AGO2 protein (black circle in the graphs) are notably different in the seven tissues. Edge thickness corresponds to the reliability score of the PPI interaction.

(a) Skin **(b)** Brain

(c) Uterine cervix **(d)** Liver

(e) Colon **(f)** Lung

(g) Kidney

Fig. 7.2 The subcellular distribution of the interacting proteins of AGO2 across tissues. The pie charts show the proportions of SCLs of interacting proteins of AGO2 protein.

Table 7.2 The SCL annotations of AGO2 protein.

SCL	Literature support	Assigned by
Nucleus	Zhang et al. [197]	BHF-UCL
Nucleoplasm (*part of* Nucleus)		HPA
Cytoplasm	Hu et al. [198], Meister et al. [199],	BHF-UCL, MGI
Cytosol (*in* Cytoplasm)		Reactome [40]
P-body (*in* Cytoplasm)	Blake et al. [24], Loedige et al. [200], Phalora et al. [201], James et al. [202],	MGI, UniProt
Polysome	Höck et al. [26]	UniProt
Cell junction		HPA
Extracellular exosome (Extracelluar space)	Beltrami et al. [23]	BHF-UCL
mRNA cap binding complex	Ryu et al. [203], Kiriakidou et al. [204]	Parkinson's UK-UCL, UniProt
RISC	James et al. [202], Yoda et al. [205], Maida et al. [206], Meister et al. [199]	BHF-UCL, MGI
RISC-loading complex	Lee et al. [207], Yoda et al. [205], Wang et al. [208], MacRae et al. [209], Robb and Rana [210], Chendrimada et al. [211]	UniProt, BHF-UCL
Intracellular ribonucleo protein complex	Höck et al. [26]	UniProt
Micro-ribonucleo protein complex	Höck et al. [26], Wakiyama et al. [212], Robb and Rana [210],	UniProt

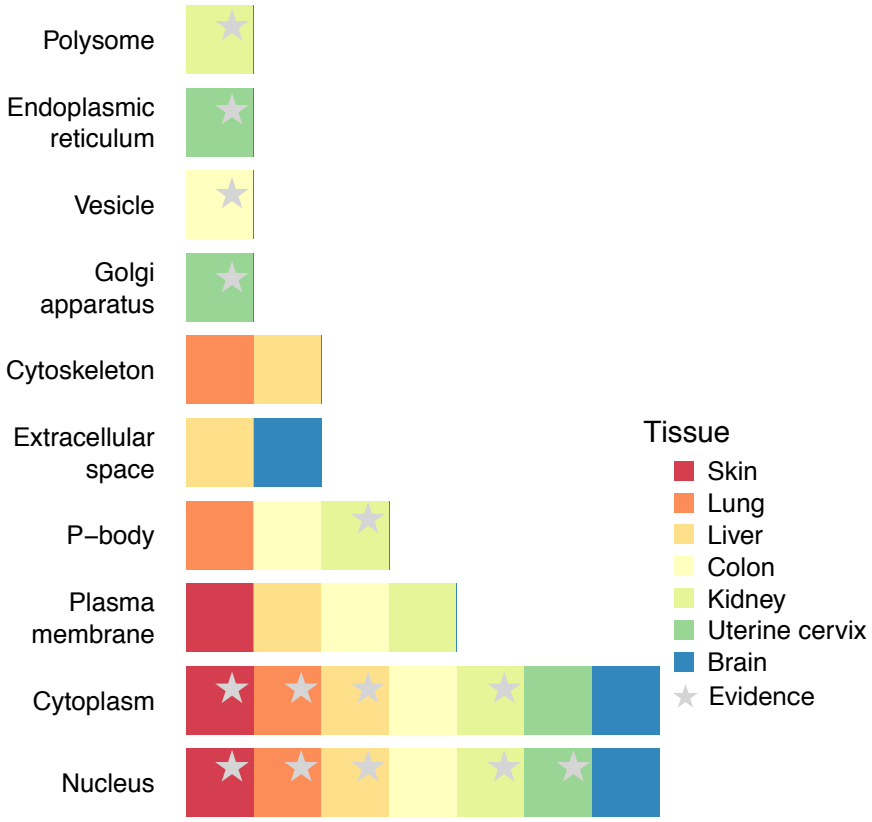


Fig. 7.3 The predicting subcellular distribution of human AGO2 across tissues.

Chapter 8

Conclusion and discussion

8.1 Conclusion

One essential task in the proteomics analysis is to explore the function of proteins in conducting and regulating the activities at the subcellular level [1]. Protein functional activities highly correspond with their subcellular distribution and molecular complexing interactions [2]. The translocalization, multi-localization, and mislocalization of protein are essential for understanding the protein functions and the mechanism of the cell. The subcellular distribution of proteins is dynamic and tissue-specific that proteins may have different roles at different SCLs in different tissue which adds another dimension of complexity in cell biology, which is the focus of this work.

Scientists have made extensive efforts to develop programs to predict the SCL of proteins. Numerous algorithms and tools have been developed in this field, based on various biological concepts and machine learning methods. Although few studies have made efforts to disease-related SCL prediction and stress derived SCL prediction, none of the existing techniques have addressed and be used for tissue-specific SCL prediction.

The first contribution of this work is the development of a new algorithm BCMRFs which improve the general multi-SCL prediction, see Chapter 4. This is accomplished by building the weighted MRFs based on the PPI network and then performing SCL label propagation to predict the SCLs of unknown proteins. We performed comprehensive experiments to evaluate the performance of human protein SCL datasets. The transductive learning from the re-balanced MLD proved to be more efficient to assign SCLs correctly. Owing to the collective MRFs which connect the binary MRFs by their spatial adjacency among SCLs, our method can achieve higher performance for predicting the multi-SCLs comparing with the state-of-the-art methods of DC-*k*NN and Hum-mPLoc 3.0. A published version of this work task can be found in Zhu and Ester [226].

The second major achievement in this work is that the tissue-specific SCLs prediction is addressed the first time. Instead of using generic physical PPINs, the tissue-specific PPINs were applied to the BCMRFs algorithm. The evaluated results demonstrated the strength of this approach on predicting tissue-specific SCLs. We successfully identified 1314 proteins which were previously proven cell line dependent on subcellular level (see Chapter 5).

Another significant contribution of this work is the development of a text mining system for mining tissue-specific SCL, see Chapter 6. It was shown the success of our dictionary-based approach on the extraction of <tissue, protein, SCL> association from PubMed abstracts. The evaluation results showed that our method achieves good precision and recall. All the data are accessible from a TS database web server via different types of search.

Finally, the application case of analysis of tissue-specific SCL of human AGO2 protein demonstrated how to use our approaches to study the dynamic SCLs of a protein of interest in different tissues and cell types, see Chapter 7. The prediction results using TS-BCMRFs showed the high accuracy in comparison with the SCL annotations from knowledge-based curation databases and manually curated tissue-specific SCL. The TS Database provides the relevant tissue-specific SCL with literature supports, and potentially complete the tissue-specific BCMRFs approach.

8.2 Discussion

The physical PPI datasets in most of the PPI databases such as IntAct [42] and BioGRID [43] are reported from different studies and techniques with huge diversity. These databases do not differentiate PPIs according to their biological contexts whereas, during different biological processes, one protein can play various roles and functions by interacting with different target proteins. When using PPINs to infer the SCLs of proteins by propagating the PPIs, the context-specific distinction of PPIs becomes crucial.

Impact on context-specific prediction.

The interactions which only occur in a specific context should not be used for the prediction of any other context. In this thesis, we focused on tissue-specific SCL of human protein which were previous proven. In Chapter 5 and Chapter 7, we performed tissue-specific SCL prediction using only the tissue-specific PPIN for the corresponding tissue. The predicted SCLs showed the significant difference from one tissue to another, which meets our expectation that the subcellular distribution of protein are tissue dependent.

Impact on generic prediction.

In the comparison of context-specific PPIN, the generic PPIN can be considered as the union of many context-specific PPINs which are connected by the proteins and interactions occurring in multiple contexts, e.g., the universal expressed proteins. However, many context-specific PPIs are exclusive to each other. When propagating the generic PPIN, the analyses can be misdirected by these 'false connections' and produces an inaccurate result. The analyses rely on context-specific PPINs should be more precise. Moreover, the union of results from the context-specific analyses supposes to be more accurate in comparison with the analyses based on the generic PPIN. It is necessary to mention that tissue-specificity which is the focus of this study, is just one aspect of the concept of context-specificity. It exists disease-specificity, condition-specificity, time-series-specificity and many other contexts which have been studied or need to be addressed in the further. Moreover, in this thesis, we mainly focus on a set of human tissues to demonstrate the strength of our approaches. Therefore, a comparison between the predicted result of a set of tissue with the generic one would be biased and not representative. Nevertheless, in Chapter 7, the predicted tissue-specific SCLs of human AGO2 protein reached high accuracy in comparison with the generic SCL annotations.

The lack of ground truth data of context-specific protein SCLs.

Although the gene/protein tissue-specificity are intensively studied, until this work was accomplished, there was not yet any tissue-specific SCL dataset available. Therefore, the collection of ground truth data and the corresponding evaluation are a challenge of this study. In Chapter 5, the SCL data detected in cell lines which were derived from corresponding tissues were used. We expect that these cell line specific SCL data reflect some aspects of tissue-specific characters. However, the quality of these datasets is limited. First of all, these cell lines are chosen for SCL experiments because of the easier manipulation procedure and the better results, especially for U2-OS cells [28]. Also, some of the cell lines are cancerous cell lines, and some are immortalized cell lines. The tissue-specific functional associations were also generated based on various types of datasets using both tissues and cell lines. Hence, we did not distinguish between healthy and diseased tissues in our study. In general, cell lines cannot represent all the tissue-specific features due to down-regulation of tissue-enriched genes [8].

It is highly recommended that researchers specify all the biological context details where their experimentally detected protein SCLs and PPIs occurred in, and this information should

be differentiated in the public databases such as UniProt, HPDB, which would make it easier to model and understand context-related phenotypes.

8.3 Future work

Recent advances in the characterization of the SCL of proteins now indicate that dynamic trafficking of multitudinous proteins over many SCLs is a central mechanism in cellular function. The major achievement of this work is the success of computational prediction of tissue-specific SCL of human proteins. In the next several years there will be significant improvements in methodology and collection of massive amounts of global data on the dynamics of protein SCL in cells under all kinds of biological context. This section discusses the further directions of dynamic protein SCL prediction standing on this work.

Extension of text mining on full text

In Chapter 6, the proposed text mining system only perform on the information extraction on the abstracts of literature. However, the full text or supplementary material, which have seemed to be an important source of tissue, cell line, and protein SCL information. Therefore, an extension to full text is one the future works which require more advanced systems to overcome the additional noise in the full text and tables.

Dynamic protein SCLs in time-series

Biological processes are often dynamic, and time sensitive. Therefore, researchers monitor the dynamic activity in a time-series manner. The time-series gene expression data were used to identify the complete set of activated genes in a biological process, to infer their rates of change, their order and their causal effects and to model dynamic systems in the cell [227, 11]. The dynamic changes of protein SCL in a time-series can accurately capture the trafficking of the protein, and thus understand the fundamental role of protein in time regulation, e.g., circadian rhythms. With the experience of tissue-specific SCL prediction, the BCMRFs algorithm and workflow in Chapter 5 can be applied on dynamic time-series PPINs to predict the SCL changes of proteins over time. However, the collection of ground truth dataset reminds as a challenge.

Isoform-specificity in SCL prediction

Some cell biology studies demonstrated functionally relevant subcellular translocation of proteins could be achieved by protein isoforms [228, 134, 229]. Two isoforms of pyruvate

kinases, PKM1 and PKM2, have diverse involvement in metabolic pathways. The cytoplasmic PKM2 isoform is essential in tumor progression. The translocation of PKM2 into the nucleus as a response to different apoptotic stimuli and PKM2 also participate in nuclear transcription complexes in response to hypoxia.

It was also suggested that a considerable part of tissue-specificity is likely to be achieved by alternative splicing and interactions involving protein isoforms [9, 230]. While the PKM2 is often considered as the embryonic isoform, the M1 isoform is expressed in differentiated cells that are actively dependent upon a high rate of energy regeneration, such as muscle and brain [231, 232].

A recent study used an image-based approach for analyzing the isoform-specificity SCL of several proteins [228]. Therefore, isoform-specificity should be an essential aspect in developing computational SCL prediction approach in the future.

Extract meaningful PPI

One benefit of understanding protein SCLs is to help to discover the function of the protein. When we use PPINs for SCL prediction, the physical contact considered in PPIN should be specific, such as tissue-specificity, time-series-specificity, disease-specificity which we have already studied or discussed in this work. However, whether the interactions that a protein experiences when it is being made (translation), folded, quality checked (post-translational modification), or degraded (post degradation) should be considered as specific and meaningful is still in doubt. For example, all proteins at one point “touch” the ribosome, many touch chaperones, and most make contact with the degradation machinery [36]. This is a critical aspect of protein function assignment issue for which the protein SCL serve as well.

References

[1] Michal Breker and Maya Schuldiner. The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nature reviews. Molecular cell biology*, 15(7):453–64, 2014. ISSN 1471-0080. doi: 10.1038/nrm3821.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, 4th edition*. Garland Science, 2002. ISBN 0815332181. doi: 10.3389/fimmu.2015.00171.

[3] Daniel V Veres, Dávid M Gyurkó, Benedek Thaler, Kristóf Z Szalay, Dávid Fazekas, Tamás Korcsmáros, and Peter Csermely. ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic acids research*, 43 (Database issue):D485–93, jan 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1007.

[4] Yuhui Hu, Hans Lehrach, and Michal Janitz. Comparative analysis of an experimental subcellular protein localization assay and in silico prediction methods. *Journal of Molecular Histology*, 40:343–352, 2009. ISSN 15672379. doi: 10.1007/s10735-009-9247-9.

[5] Mien-Chie M.-C. Hung and Wolfgang Link. Protein localization in disease and therapy. *Journal of cell science*, 124(Pt 20):3381–92, oct 2011. ISSN 1477-9137. doi: 10.1242/jcs.089110.

[6] Constance J. Jeffery. Why study moonlighting proteins? *Frontiers in Genetics*, 6:211, jun 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00211.

[7] Michael J. Rust, Joseph S. Markson, William S. Lane, Daniel S. Fisher, and Erin K. O'Shea. Ordered phosphorylation governs oscillation of a three-protein circadian clock. *Science*, 318(5851):809–812, 2007. ISSN 00368075. doi: 10.1126/science.1148596.

[8] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-henrik P.-H. Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle Von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar Von Heijne, Jens Nielsen, Fredrik Pontén, M. Uhlen, Linn Fagerberg, B. M. Hallstrom, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, A. Sivertsson, Caroline Kampf, E. Sjostedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic,

Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-henrik P.-H. Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, K. von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, G. von Heijne, Jens Nielsen, F. Ponten, Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-henrik P.-H. Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle Von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar Von Heijne, Jens Nielsen, Fredrik Pontén, M. Uhlen, Linn Fagerberg, B. M. Hallstrom, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, A. Sivertsson, Caroline Kampf, E. Sjostedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-henrik P.-H. Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, K. von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, G. von Heijne, Jens Nielsen, and F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419, jan 2015. ISSN 0036-8075. doi: 10.1126/science.1260419.

[9] Esti Yeger-Lotem and Roded Sharan. Human protein interaction networks across tissues and diseases. *Frontiers in genetics*, 6:257, 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00257.

[10] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, Daniel I Chasman, Garret A FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–76, 2015. ISSN 1546-1718. doi: 10.1038/ng.3259.

[11] Thomas Wallach, Katja Schellenberg, Bert Maier, Ravi Kiran Reddy Kalathur, Pablo Porras, Erich E. Wanker, Matthias E. Futschik, and Achim Kramer. Dynamic Circadian Protein-Protein Interaction Networks Predict Temporal Organization of Cellular Functions. *PLoS Genetics*, 9(3):e1003398, mar 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003398.

[12] KiYoung Lee, Min-Kyung Sung, Jihyun Kim, Kyung Kim, Junghyun Byun, Hyojung Paik, Bongkeun Kim, Won-Ki Huh, and Trey Ideker. Proteome-wide remodeling of protein location and function by stress. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30):E3157–66, jul 2014. ISSN 1091-6490. doi: 10.1073/pnas.1318881111.

[13] Tobias B. Dansen and Boudewijn M.T. Burgering. Unravelling the tumor-suppressive functions of FOXO proteins, 2008. ISSN 09628924.

[14] Gorbatchev Ambroise, Alain Portier, Nathalie Roders, Damien Arnoult, and Aimé Vazquez. Subcellular localization of PUMA regulates its pro-apoptotic activity in

Burkitt's lymphoma B cells. *Oncotarget*, 6(35):38181–94, nov 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.5901.

[15] Gideon D. Matthews, Noa Gur, Werner J. H. Koopman, Ophry Pines, and Lily Vardimon. Weak mitochondrial targeting sequence determines tissue-specific subcellular localization of glutamine synthetase in liver and brain cells. *Journal of cell science*, 123(Pt 3):351–359, 2010. ISSN 0021-9533. doi: 10.1242/jcs.060749.

[16] A R Cho, K J Yang, Y Bae, Y Y Bahk, E Kim, H Lee, J K Kim, W Park, H Rhim, S Y Choi, T Imanaka, S Moon, J Yoon, and S K Yoon. Tissue-specific expression and subcellular localization of ALADIN, the absence of which causes human triple A syndrome. *Exp Mol Med*, 41(6):381–386, 2009. ISSN 1226-3613. doi: 10.3858/emm.2009.41.6.043[doi].

[17] Nishi R Sharma, Xiaohong Wang, Vladimir Majerciak, Masahiko Ajiro, Michael Kruhlak, Craig Meyers, and Zhi-Ming Zheng. Cell Type- and Tissue Context-dependent Nuclear Distribution of Human Ago2. *The Journal of biological chemistry*, 291(5):2302–9, jan 2016. ISSN 1083-351X. doi: 10.1074/jbc.C115.695049.

[18] Marilyn E Thompson. BRCA1 16 years later: nuclear import and export processes. *The FEBS journal*, 277(15):3072–8, 2010. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2010.07733.x.

[19] Eleni Mylona, Savvas Melissaris, Alexandros Nomikos, Irene Theohari, Ioanna Giannopoulou, Konstantinos Tzelepis, and Lydia Nakopoulou. Effect of BRCA1 immunohistochemical localizations on prognosis of patients with sporadic breast carcinomas. *Pathology - Research and Practice*, 210(8):533–540, aug 2014. ISSN 03440338. doi: 10.1016/j.prp.2014.05.009.

[20] Sébastien Rey, Michael Acab, Jennifer L Gardy, Matthew R Laird, Katalin DeFays, Christophe Lambert, and Fiona S L Brinkman. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic acids research*, 33(Database issue):D164–8, jan 2005. ISSN 1362-4962. doi: 10.1093/nar/gki027.

[21] Chang Jin Shin, Simon Wong, Melissa J Davis, and Mark A Ragan. Protein-protein interaction as a predictor of subcellular location. *BMC systems biology*, 3:28, jan 2009. ISSN 1752-0509. doi: 10.1186/1752-0509-3-28.

[22] Julia Höck and Gunter Meister. The Argonaute protein family. *Genome biology*, 9(2):210, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-2-210.

[23] Cristina Beltrami, Marie Besnier, Saran Shantikumar, Andrew I.U. Shearn, Cha Rajakaruna, Abas Laftah, Fausto Sessa, Gaia Spinetti, Enrico Petretto, Gianni D. Angelini, and Costanza Emanueli. Human Pericardial Fluid Contains Exosomes Enriched with Cardiovascular-Expressed MicroRNAs and Promotes Therapeutic Angiogenesis. *Molecular Therapy*, 25(3):679–693, 2017. ISSN 15250024. doi: 10.1016/j.ymthe.2016.12.022.

[24] Judith A. Blake, Janan T. Eppig, James A. Kadin, Joel E. Richardson, Cynthia L. Smith, Carol J. Bult, A. Anagnostopoulos, R. M. Baldarelli, J. S. Beal, S. M. Bello, O. Blodgett, N. E. Butler, L. E. Corbani, H. Dene, H. J. Drabkin, K. L. Forthofer,

S. L. Giannatto, P. Hale, D. P. Hill, L. Hutchins, M. Knowlton, A. Lavertu, M. Law, J. R. Lewis, V. Lopez, D. Maghini, D. Perry, M. McAndrews, D. Miers, H. Montenko, L. Ni, H. Onda, J. M. Recla, D. J. Reed, B. Richards-Smith, D. Sitnikov, M. Tomczuk, L. Wilming, and Y. Zhu. Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, 45(D1):D723–D729, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1040.

[25] Keith T. Gagnon, Liande Li, Yongjun Chu, Bethany A. Janowski, and David R. Corey. RNAi factors are present and active in human cell nuclei. *Cell Reports*, 6(1):211–221, jan 2014. ISSN 22111247. doi: 10.1016/j.celrep.2013.12.013.

[26] Julia Höck, Lasse Weinmann, Christine Ender, Sabine Rüdel, Elisabeth Kremmer, Monika Raabe, Henning Urlaub, and Gunter Meister. Proteomic and functional analysis of Argonaute-containing mRNA-protein complexes in human cells. *EMBO Reports*, 8(11):1052–1060, 2007. ISSN 1469221X. doi: 10.1038/sj.embor.7401088.

[27] Yu-Ling Shih and Lawrence Rothfield. The bacterial cytoskeleton. *Microbiology and molecular biology reviews : MMBR*, 70(3):729–54, sep 2006. ISSN 1092-2172. doi: 10.1128/MMBR.00017-06.

[28] Peter J. Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M. Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M. Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M. Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P. Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S. Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, may 2017. ISSN 0036-8075. doi: 10.1126/science.aal3321.

[29] Agnieszka Chacinska, Carla M. Koehler, Dusanka Milenkovic, Trevor Lithgow, and Nikolaus Pfanner. Importing Mitochondrial Proteins: Machineries and Mechanisms, 2009. ISSN 00928674.

[30] Oliver Schmidt, Nikolaus Pfanner, and Chris Meisinger. Mitochondrial protein import: From proteomics to functional mechanisms, 2010. ISSN 14710072.

[31] William Wickner and Randy Schekman. Protein translocation across biological membranes. *Science (New York, N.Y.)*, 310(5753):1452–6, 2005. ISSN 1095-9203. doi: 10.1126/science.1113752.

[32] Allison Lange, Ryan E. Mills, Christopher J. Lange, Murray Stewart, Scott E. Devine, and Anita H. Corbett. Classical nuclear localization signals: Definition, function, and interaction with importin α, 2007. ISSN 00219258.

[33] L I Ashmarina, A V Pshezhetsky, S S Branda, G Isaya, and G A Mitchell. 3-Hydroxy-3-methylglutaryl coenzyme A lyase: targeting and processing in peroxisomes and mitochondria. *Journal of lipid research*, 40(1):70–5, jan 1999. ISSN 0022-2275.

[34] Shao-Chun Wang and Mien-Chie Hung. Nuclear translocation of the epidermal growth factor receptor family membrane tyrosine kinase receptors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15 (21):6484–6489, 2009. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-08-2813.

[35] Antonella Caccamo, Smita Majumder, Janice J. Deng, Yidong Bai, Fiona B. Thornton, and Salvatore Oddo. Rapamycin rescues TDP-43 mislocalization and the associated low molecular mass neurofilament instability. *Journal of Biological Chemistry*, 284 (40):27416–27424, 2009. ISSN 00219258. doi: 10.1074/jbc.M109.031278.

[36] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, jun 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000807.

[37] G Y Wiederschain. Protein-protein interactions. A molecular cloning manual. In *Biochemistry*, volume 71, page 2979. 2006. ISBN 9780879697228. doi: 10.1134/S0006297906060162.

[38] Saliha Ece Acuner Ozbabacan, Hatice Billur Engin, Attila Gursoy, and Ozlem Keskin. Transient proteinprotein interactions, 2011. ISSN 17410126.

[39] 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. I. 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(November 2008): 767–772, 2009. ISSN 03051048. doi: 10.1093/nar/gkn892.

[40] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1132.

[41] Ioannis Xenarios, Lukasz Salwínski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1): 303–5, jan 2002. ISSN 1362-4962.

[42] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum,

Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue): D358–63, jan 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1115.

[43] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1102.

[44] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(Database issue):D857–61, jan 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr930.

[45] Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona Brinkman, Gianni Cesareni, Andrew Chatr-aryamontri, Emilie Chautard, Carol Chen, Marine Dumousseau, Johannes Goll, Robert Hancock, Linda I Hannick, Igor Jurisica, Jyoti Khadake, David J Lynn, Usha Mahadevan, Livia Perfetto, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Lukasz Salwinski, Volker Stümpflen, Mike Tyers, Peter Uetz, Ioannis Xenarios, and Henning Hermjakob. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(6):626–626, 2012. ISSN 1548-7091. doi: 10.1038/nmeth0612-626a.

[46] Eric W. Deutsch, Sandra Orchard, Pierre-Alain Binz, Wout Bittremieux, Martin Eisenacher, Henning Hermjakob, Shin Kawano, Henry Lam, Gerhard Mayer, Gerben Menschaert, Yasset Perez-Riverol, Reza M. Salek, David L. Tabb, Stefan Tenzer, Juan Antonio Vizcaíno, Mathias Walzer, and Andrew R. Jones. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *Journal of Proteome Research*, 16(12):4288–4298, dec 2017. ISSN 1535-3893. doi: 10.1021/acs.jproteome.7b00370.

[47] Einat Sprinzak, Shmuel Sattath, and Hanah Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003. ISSN 00222836. doi: 10.1016/S0022-2836(03)00239-0.

[48] Bruno Aranda, Hagen Blankenburg, Samuel Kerrien, Fiona S L Brinkman, Arnaud Ceol, Emilie Chautard, Jose M Dana, Javier De Las Rivas, Marine Dumousseau, Eugenia Galeota, Anna Gaulton, Johannes Goll, Robert E W Hancock, Ruth Isserlin, Rafael C Jimenez, Jules Kerssemakers, Jyoti Khadake, David J Lynn, Magali Michaut, Gavin O'Kelly, Keiichiro Ono, Sandra Orchard, Carlos Prieto, Sabry Razick, Olga Rigina, Lukasz Salwinski, Milan Simonovic, Sameer Velankar, Andrew Winter, Guanming Wu, Gary D Bader, Gianni Cesareni, Ian M Donaldson, David Eisenberg, Gerard J Kleywegt, John Overington, Sylvie Ricard-Blum, Mike Tyers, Mario Albrecht, and Henning Hermjakob. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature methods*, 8(7):528–9, jun 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1637.

[49] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. mentha: a resource for browsing integrated protein-interaction networks. *Nature methods*, 10(8):690–1, aug 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2561.

[50] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, Michael Kuhn, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(Database issue):D447–52, jan 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1003.

[51] Paul Shannon, Andrew Markiel, 2 Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13):2498–2504, 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303. metabolite.

[52] BJÖRN Sommer, BENJAMIN Kormeier, Pavel S. DEMENKOV, PATRIZIO Arrigo, KLAUS Hippe, ÖZGÜR Ates, Alexey V. KOCHETOV, Vladimir A. IVANISENKO, NIKOLAY A. Kolchanov, and Ralf HOFESTÄDT. Subcellular localization charts: a new visual methodology for the semi-automatic localization of protein-related data sets. *Journal of bioinformatics and computational biology*, 11(1):1340005, feb 2013. ISSN 0219-7200. doi: 10.1142/S0219720013400052.

[53] Christoph Brinkrolf, Sebastian Jan Janowski, Benjamin Kormeier, Martin Lewinski, Klaus Hippe, Daniela Borck, and Ralf Hofestädt. VANESA - a software application for the visualization and analysis of networks in system biology applications. *Journal of integrative bioinformatics*, 11(2):239, jan 2014. ISSN 1613-4516. doi: 10.2390/biecoll-jib-2014-239.

[54] Georgios A Pavlopoulos, Dimitris Malliarakis, Nikolas Papanikolaou, Theodosis Theodosiou, Anton J Enright, and Ioannis Iliopoulos. Visualizing genome and systems biology : technologies , tools , implementation techniques and trends , past , present and future. *GigaScience*, 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0077-2.

[55] Elise A R Serin, Harm Nijveen, Henk W M Hilhorst, and Wilco Ligterink. Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in plant science*, 7:444, 2016. ISSN 1664-462X. doi: 10.3389/fpls.2016.00444.

[56] Xiaoli; Brooks, Steve; Gelman, Andrew; Jones, Galin; Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. ISBN 9781420079418.

[57] Ilker Yildirim. Bayesian Inference: Gibbs Sampling. 2012.

[58] George Casella and Edward I. George. Explaining the gibbs sampler. *American Statistician*, 46(3):167–174, 1992. ISSN 15372731. doi: 10.1080/00031305.1992.10475878.

[59] Julian Besag. On the Statistical Analisys of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3):259–302, 1986. ISSN 00359246.

[60] Zhi Wei and Hongzhe Li. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, jun 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm129.

[61] Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Marco C. A. M. Bink, Roeland C. H. J. van Ham, and Cajo J. F. ter Braak. Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. *PLoS ONE*, 5(2):e9293, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009293.

[62] Daphne. Koller and Nir. Friedman. *Probabilistic graphical models : principles and techniques*. MIT Press, 2009. ISBN 0262013193.

[63] Pushmeet Kohli, M. Pawan Kumar, and Philip H.S. Torr. P3 and beyond: Solving energies with higher order cliques. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007. ISBN 1424411807. doi: 10.1109/CVPR.2007.383204.

[64] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms. In *Hybrid Artificial Intelligence Systems*, pages 110–121. Springer, Cham, 2014. doi: 10.1007/978-3-319-07617-1_10.

[65] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:1–14, 2015. ISSN 09252312. doi: 10.1016/j.neucom.2014.08.091.

[66] Ronen Feldman and James Sanger. *The Text Mining Handbook Text*. Number 4. Cambridge University Press, 2011. ISBN 978-0521836579. doi: 10.1088/1751-8113/44/8/085201.

[67] Yu Zhang, Mengdong Chen, and Lianzhong Liu. A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 681–685. IEEE, sep 2015. ISBN 978-1-4799-8352-0. doi: 10.1109/ICSESS.2015.7339149.

[68] Dietrich Rebholz-Schuhmann, Anika Oellrich, and Robert Hoehndorf. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, 13(12):829–839, nov 2012. ISSN 1471-0056. doi: 10.1038/nrg3337.

[69] Wilco W.M. Fleuren and Wynand Alkema. Application of text mining in the biomedical domain, 2015. ISSN 10959130.

[70] Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472, jun 2013. ISSN 0268-4012. doi: 10.1016/J.IJINFOMGT.2013.01.001.

[71] Yoko Kobayashi and Kazuhiko Tsuda. Customer satisfaction factor extraction method using text mining. In *2017 International Conference on Emerging Trends and Innovation in ICT (ICEI)*, pages 86–91. IEEE, feb 2017. ISBN 978-1-5090-3404-8. doi: 10.1109/ETIICT.2017.7977016.

[72] Kenneth A Myers and Christopher Janetopoulos. Recent advances in imaging subcellular processes. *F1000Research*, 5, 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8399.1.

[73] Trisha N. Davis. Protein localization in proteomics, 2004. ISSN 13675931.

[74] Andy Christoforou, Claire M. Mulvey, Lisa M. Breckels, Aikaterini Geladaki, Tracey Hurrell, Penelope C. Hayward, Thomas Naake, Laurent Gatto, Rosa Viner, Alfonso Martinez Arias, and Kathryn S. Lilley. A draft map of the mouse pluripotent stem cell spatial proteome. *Nature Communications*, 7:9992, jan 2016. ISSN 2041-1723. doi: 10.1038/ncomms9992.

[75] Daniel N Itzhak, Stefka Tyanova, Jürgen Cox, and Georg HH Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife*, 5, jun 2016. ISSN 2050-084X. doi: 10.7554/eLife.16950.

[76] Song Yi Lee, Myeong Gyun Kang, Jong Seok Park, Geunsik Lee, Alice Y. Ting, and Hyun Woo Rhee. APEX Fingerprinting Reveals the Subcellular Localization of Proteins of Interest. *Cell Reports*, 15(8):1837–1847, may 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.04.064.

[77] Kyle J. Roux, Dae In Kim, Manfred Raida, and Brian Burke. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of Cell Biology*, 196(6):801–810, mar 2012. ISSN 0021-9525. doi: 10.1083/jcb.201112098.

[78] Charlotte Stadler, Elton Rexhepaj, Vasanth R Singan, Robert F Murphy, Rainer Pepperkok, Mathias Uhlén, Jeremy C Simpson, and Emma Lundberg. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nature Methods*, 10(4):315–323, apr 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2377.

[79] Shruti Rastogi and Burkhard Rost. LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. *Nucleic acids research*, 39(Database issue):D230–4, jan 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq927.

[80] Eurie L. Hong, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Venkat S. Malladi, J. Seth Strattan, Benjamin C. Hitz, Idan Gabdank, Aditi K. Narayanan, Marcus Ho, Brian T. Lee, Laurence D. Rowe, Timothy R. Dreszer, Greg R. Roe, Nikhil R. Podduturi, Forrest Tanaka, Jason A. Hilton, and J. Michael Cherry. Principles of metadata organization at the ENCODE data coordination center. *Database*, 2016: baw001, mar 2016. ISSN 1758-0463. doi: 10.1093/database/baw001.

[81] 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): D767–72, jan 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn892.

[82] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1108.

[83] Jennifer L Gardy and Fiona S L Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature reviews. Microbiology*, 4(10):741–51, oct 2006. ISSN 1740-1534. doi: 10.1038/nrmicro1494.

[84] Torsten Blum, Sebastian Briesemeister, Oliver Kohlbacher, O Emanuelsson, S Brunak, G von Heijne, H Nielson, R Nair, B Rost, O Emanuelsson, H Nielson, G von Heijne, JD Bendtsen, H Nielsen, G von Heijne, S Brunak, O Emanuelsson, H Nielson, S Brunak, G von Heijne, H Bannai, Y Tamada, O Maruyama, K Nakai, S Miyano, EI Petsalaki, PG Bagos, ZI Litou, SJ Hamodrakas, Y Fujiwara, M Asogawa, M Boden, J Hawkins, I Small, N Peeters, F Legeai, C Lurin, M Cokol, R Nair, B Rost, MA Andrade, SI O'Donoghue, B Rost, J Cedano, P Aloy, JA Pérez-Pons, E Querol, A Reinhardt, T Hubbard, S Hua, Z Sun, KJ Park, M Kanehisa, D Xie, A Li, M Wang, Z Fan, H Feng, J Guo, Y Lin, A Pierleoni, PL Martelli, PL Fariselli, R Casadio, Q Cui, T Jiang, B Liu, S Ma, KC Chou, Y Cai, P Horton, KJ Park, T Obayashi, N Fujita, H Harada, CJ Adams-Collier, K Nakai, A Höglund, P Dönnes, Torsten Blum, HW Adolph, Oliver Kohlbacher, KC Chou, Y Cai, MS Scott, DY Thomas, MT Hallett, R Nair, B Rost, Z Lu, D Szafron, R Greiner, P Lu, DS Wishart, B Poulin, J Anvik, C Macdonell, R Eisner, Z Lei, Y Dai, WL Huanq, CW Tunq, SY SW Ho, SF Hwang, SY SW Ho, S Brady, H Shatkay, A Fyshe, Y Liu, D Szafron, R Greiner, P Lu, R Nair, B Rost, H Shatkay, A Höglund, S Brady, Torsten Blum, P Dönnes, Oliver Kohlbacher, C Guda, S Subramaniam, M Bhasin, GP Raghava, KC Chou, HB Shen, KC Chou, HB Shen, YO Shen, G Burger, J Liu, S Kang, G Tang, LBM Ellis, T Li, KC Chou, HB Shen, HB Shen, KC Chou, EM Marcotte, I Xenarios, AMvan der Bliek, D Eisenberg, M Pellegrini, EM Marcotte, MJ Thompson, D Eisenberg, TO Yeates, K Enault, C Suhre, O Poirot, JM Clavarie, EM Marcotte, null Ashburner, KC Chou, Y Cai, Z Lu, L Hunter, null Bairoch, HB Shen, J Yanq, KC Chou, EM Zdobnov, R Apweiler, KC Chou, Y Cai, NJ Mulder, R Casadio, PL Martelli, A Pierleoni, VN Vapnik, CC Chang, and CJ Lin. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, 10(1):274, jan 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-274.

[85] Alessandro Adelfio, Viola Volpato, and Gianluca Pollastri. SCLpredT: Ab initio and homology-based prediction of subcellular localization by N-to-1 neural networks. *SpringerPlus*, 2:502, 2013. ISSN 2193-1801. doi: 10.1186/2193-1801-2-502.

[86] Shibiao Wan, Man Wai Mak, and Sun Yuan Kung. GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *Journal of Theoretical Biology*, 323:40–48, 2013. ISSN 00225193. doi: 10.1016/j.jtbi.2013.01.012.

[87] Kenta Nakai. Protein sorting signals and prediction of subcellular localization. *Advances in protein chemistry*, 54:277–344, 2000. ISSN 0065-3233. doi: 10.1016/S0065-3233(00)54009-1.

[88] Markus Brameier, Andrea Krings, and Robert M MacCallum. NucPred–predicting nuclear localization of proteins. *Bioinformatics (Oxford, England)*, 23(9):1159–60, 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm066.

[89] Olof Emanuelsson, Henrik Nielsen, and Gunnar Von Heijne. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, 8(5):978–984, 1999. ISSN 09618368. doi: 10.1110/ps.8.5.978.

[90] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx431.

[91] Paul Horton, Keun Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C J Adams-Collier, and Kenta Nakai. WoLF PSORT: Protein localization predictor. *Nucleic Acids Research*, 35(SUPPL.2):585–587, 2007. ISSN 03051048. doi: 10.1093/nar/gkm259.

[92] Xiaotong Guo, Fulin Liu, Ying Ju, Zhen Wang, and Chunyu Wang. Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier. *Scientific Reports*, 6(February):28087, 2016. ISSN 2045-2322. doi: 10.1038/srep28087.

[93] Ying-Li Chen and Qian-Zhong Li. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. *Journal of Theoretical Biology*, 248(2):377–381, 2007. ISSN 00225193. doi: 10.1016/j.jtbi.2007.05.019.

[94] Kuo Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function and Genetics*, 43(3):246–255, 2001. ISSN 08873585. doi: 10.1002/prot.1035.

[95] Evangelia I. Petsalaki, Pantelis G. Bagos, Zoi I. Litou, and Stavros J. Hamodrakas. PredSL: A tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics, Proteomics and Bioinformatics*, 4(1):48–55, 2006. ISSN 16720229. doi: 10.1016/S1672-0229(06)60016-8.

[96] O Emanuelsson, H Nielsen, S Brunak, and G von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016, 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.3903.

[97] Hang Zhou, Yang Yang, Hong Bin Shen, and John Hancock. Hum-mPLoc 3.0: Prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics*, 33(6):843–853, 2017. ISSN 14602059. doi: 10.1093/bioinformatics/btw723.

[98] Kai Wang, Le-Le Hu, Xiao-He Shi, Ying-Song Dong, Hai-Peng Li, and Tie-Qiao Wen. PSCL: predicting protein subcellular localization based on optimal functional domains. *Protein and peptide letters*, 19(1):15–22, 2012. ISSN 0929-8665. doi: 10.2174/092986612798472820.

[99] Peilin Jia, Ziliang Qian, ZhenBin Zeng, Yudong Cai, and Yixue Li. Prediction of subcellular protein localization based on functional domain composition. *Biochemical and Biophysical Research Communications*, 357(2):366–370, 2007. ISSN 0006291X. doi: 10.1016/j.bbrc.2007.03.139.

[100] Shengnan Tang, Tonghua Li, Peisheng Cong, Wenwei Xiong, Zhiheng Wang, and Jiangming Sun. PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif. *Nucleic acids research*, 41(Web Server issue), 2013. ISSN 13624962. doi: 10.1093/nar/gkt428.

[101] Tien Ho Lin, Robert F. Murphy, and Ziv Bar-Joseph. Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):441–451, 2011. ISSN 15455963. doi: 10.1109/TCBB.2009.82.

[102] Hao Yu, Wenjian Bi, Chenxing Liu, Yanlong Zhao, Ji-Feng Zhang, Dai Zhang, and Weihua Yue. Protein-interaction-network-based analysis for genome-wide association analysis of schizophrenia in Han Chinese population. *Journal of psychiatric research*, 50:73–8, mar 2014. ISSN 1879-1379. doi: 10.1016/j.jpsychires.2013.11.014.

[103] Chi-Hua Tung, Chi-Wei Chen, Han-Hao Sun, and Yen-Wei Chu. Predicting human protein subcellular localization by heterogeneous and comprehensive approaches. *PloS one*, 12(6):e0178832, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0178832.

[104] Sebastian Briesemeister, Jörg Rahnenführer, and Oliver Kohlbacher. YLoc–an interpretable web server for predicting subcellular localization. *Nucleic acids research*, 38 (Web Server issue):W497–502, jul 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq477.

[105] Christian J A Sigrist, Edouard De Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(D1), 2013. ISSN 03051048. doi: 10.1093/nar/gks1067.

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

[107] T. K.B. Gandhi, Jun Zhong, Suresh Mathivanan, L. Karthick, K. N. Chandrika, S. Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D. Boeke, Giovanni Parmigiani, Jörg Schultz, Joel S. Bader, and Akhilesh Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, 2006. ISSN 10614036. doi: 10.1038/ng1747.

[108] Kiyoung Lee, Han-Yu Chuang, Andreas Beyer, Min-Kyung Sung, Won-Ki Huh, Bonghee Lee, and Trey Ideker. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research*, 36(20): e136, nov 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn619.

[109] Pufeng Du and Lusheng Wang. Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. *PloS one*, 9(1):e86879, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0086879.

[110] Hong-Bin Shen and Kuo-Chen Chou. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical biochemistry*, 394(2):269–74, nov 2009. ISSN 1096-0309. doi: 10.1016/j.ab.2009.07.046.

[111] Sebastian Briesemeister, Jörg Rahnenführer, Oliver Kohlbacher, J. Rahnenfuhrer, and Oliver Kohlbacher. Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics (Oxford, England)*, 26(9):1232–8, may 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq115.

[112] Jonathan Q Jiang and Maoying Wu. Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study. *BMC bioinformatics*, 13 Suppl 1(Suppl 10):S20, jan 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S10-S20.

[113] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21:302–310, 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti1054.

[114] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697–700, 2003. ISSN 1087-0156. doi: 10.1038/nbt825.

[115] Shira Mintz-Oron, Asaph Aharoni, Eytan Ruppin, and Tomer Shlomi. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics (Oxford, England)*, 25(12):i247–52, jun 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp209.

[116] Ananda Mondal and Jianjun Hu. Network Based Prediction of Protein Localisation Using Diffusion Kernel. *Int. J. Data Min. Bioinformatics*, 9(4):386–400, 2014. ISSN 17485673.

[117] KiYoung Lee, Kyunghee Byun, Wonpyo Hong, Han Yu Chuang, Chan Gi Pack, Enkhjargal Bayarsaikhan, Sun Ha Paek, Hyosil Kim, Hye Young Shin, Trey Ideker, and Bonghee Lee. Proteome-wide discovery of mislocated proteins in cancer. *Genome Research*, 23(8):1283–1294, 2013. ISSN 10889051. doi: 10.1101/gr.155499.113.

[118] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(88):88, jan 2007. ISSN 1744-4292. doi: 10.1038/msb4100129.

[119] Minghua Deng, K U I Zhang, Shipra Mehta, and Ting Chen. Prediction of Protein Function Using Protein – Protein Interaction Data. *Journal of Computational Biology*, 10(6):947–960, 2003.

[120] R. W. M. WEDDERBURN. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447, dec 1974. ISSN 0006-3444. doi: 10.1093/biomet/61.3.439.

[121] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov Random Field modeling, inference and learning in computer vision and image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, nov 2013. ISSN 10773142. doi: 10.1016/j.cviu.2013.07.004.

[122] Shantanu Godbole and Sunita Sarawagi. Discriminative Methods for Multi-labeled Classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer Berlin Heidelberg, 2004. doi: 10.1007/978-3-540-24775-3_5.

[123] Fabian L Wauthier, Michael I Jordan, and Nebojsa Jojic. Efficient Ranking from Pairwise Comparisons. In *ICML*, volume 28, pages 1–9, 2013.

[124] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010. ISSN 14337851. doi: 10.1007/978-0-387-09823-4_34.

[125] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, jul 2007. ISSN 00313203. doi: 10.1016/j.patcog.2006.12.019.

[126] Min-ling Zhang and Zhi-hua Zhou. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006. ISSN 1041-4347. doi: 10.1109/TKDE.2006.162.

[127] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008. ISSN 08856125. doi: 10.1007/s10994-008-5077-3.

[128] Xin Li and Yuhong Guo. Active learning with multi-label SVM classification. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 1479–1485, 2013. ISBN 9781577356332.

[129] Suyu Mei. Multi-label multi-Kernel transfer learning for human protein subcellular localization. *PLoS ONE*, 7(6):e37716, jun 2012. ISSN 19326203. doi: 10.1371/journal.pone.0037716.

[130] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics*, 13(1):290, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-290.

[131] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, oct 2012. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2012.03.014.

[132] Jianjun He, Hong Gu, and Wenqi Liu. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PLoS ONE*, 7(6), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0037155.

[133] Constance J. Jeffery. Multifunctional proteins: Examples of gene sharing. *Annals of Medicine*, 35(1):28–35, 2003. ISSN 07853890. doi: 10.1080/07853890310004101.

[134] Gabriella Pinto, Abdulrab Ahmed M Alhaiek, and Jasminka Godovac-Zimmermann. Proteomics reveals the importance of the dynamic redistribution of the subcellular location of proteins in breast cancer cells. *Expert Review of Proteomics*, 12(1):61–74, 2015. ISSN 1478-9450. doi: 10.1586/14789450.2015.1002474.

[135] Zhonghao Liu and Jianjun Hu. Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction 1 Introduction. *Methods*, 93:119–127, 2015. ISSN 10462023. doi: 10.1016/j.ymeth.2015.09.022.

[136] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29:569–574, 2013. doi: 10.1016/j.tig.2013.05.010.

[137] Ruth Barshir, Omer Basha, Amir Eluk, Ilan Y. Smoly, Alexander Lan, and Esti Yeger-Lotem. The TissueNet database of human tissue protein-protein interactions. *Nucleic acids research*, 41(Database issue):D841–D844, jan 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1198.

[138] Omer Basha, Ruth Barshir, Moran Sharon, Eugene Lerman, Binyamin F Kirson, Idan Hekselman, and Esti Yeger-Lotem. The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues. *Nucleic acids research*, 45(D1): D427–D431, jan 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1088.

[139] J. Song, Z. Wang, and R. M. Ewing. Integrated analysis of the Wnt responsive proteome in human cells reveals diverse and cell-type specific networks. *Mol. BioSyst.*, 10(1):45–53, 2014. ISSN 1742-206X. doi: 10.1039/C3MB70417C.

[140] Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Molecular systems biology*, 5(260):260, apr 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.17.

[141] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–50, dec 2010. ISSN 1546-1696. doi: 10.1038/nbt1210-1248.

[142] J. L. Gardy. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, 31(13):3613–3617, jul 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg602.

[143] Noah Lee, Andrew F. Laine, and R. Theodore Smith. No Title, 2009. ISSN 16800737.

[144] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(1):212–224, jan 2017. ISSN 1545-5963. doi: 10.1109/TCBB.2016.2527657.

[145] Steven E. Arnold, Bradley T. Hyman, Jill Flory, Antonio R. Damasio, and Gary W. Van Hoesen. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with alzheimer's disease. *Cerebral Cortex*, 1(1):103–116, 1991. ISSN 14602199. doi: 10.1093/cercor/1.1.103.

[146] Cajo J. F. ter Braak and Jasper A. Vrugt. Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446, oct 2008. ISSN 0960-3174. doi: 10.1007/s11222-008-9104-9.

[147] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on PAMI*, 23(11):1222–12391, 2001. URL http://www.csd.uwo.ca/$\sim$yuri/Papers/pami01.pdf.

[148] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, Christian Stolte, Seán I O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database : the journal of biological databases and curation*, 2014:bau012, jan 2014. ISSN 1758-0463. doi: 10.1093/database/bau012.

[149] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6913 LNAI, pages 145–158. Springer Berlin Heidelberg, 2011. ISBN 9783642238079. doi: 10.1007/978-3-642-23808-6_10.

[150] Nancy Y. Yu, James R. Wagner, Matthew R. Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S Cenk Sahinalp, Martin Ester, Leonard J. Foster, Fiona S L Brinkman, S. Cenk Sahinalp, Martin Ester, Leonard J. Foster, and Fiona S L Brinkman. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, jul 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq249.

[151] Edroaldo Lummertz da Rocha, Choong Yong Ung, Cordelia D. McGehee, Cristina Correia, Hu Li, Edroaldo Lummertz da Rocha, Choong Yong Ung, Cordelia D. McGehee, Cristina Correia, and Hu Li. NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Research*, 44(10), jun 2016. doi: 10.1093/nar/gkw166.

[152] The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research*, 42:D191–8, 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1140.

[153] Martin Wühr, Thomas Güttler, Leonid Peshkin, Graeme C. McAlister, Matthew Sonnett, Keisuke Ishihara, Aaron C. Groen, Marc Presler, Brian K. Erickson, Timothy J. Mitchison, Marc W. Kirschner, and Steven P. Gygi. The Nuclear Proteome

of a Vertebrate. *Current Biology*, 25(20):2663–2671, 2015. ISSN 09609822. doi: 10.1016/j.cub.2015.08.047.

[154] Luciana Barbini, Joaquin Rodríguez, Fernando Dominguez, and Felix Vega. Glyceraldehyde-3-phosphate dehydrogenase exerts different biologic activities in apoptotic and proliferating hepatocytes according to its subcellular localization. *Molecular and cellular biochemistry*, 300(1-2):19–28, jun 2007. ISSN 0300-8177. doi: 10.1007/s11010-006-9341-1.

[155] Xiaofeng Zuo, Ben Fogelgren, and Joshua H. Lipschutz. The Small GTPase Cdc42 Is Necessary for Primary Ciliogenesis in Renal Tubular Epithelial Cells. *Journal of Biological Chemistry*, 286(25), jun 2011. doi: 10.1074/jbc.M111.238469.

[156] G Lonart and T C Südhof. Region-specific phosphorylation of rabphilin in mossy fiber nerve terminals of the hippocampus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 18(2):634–40, jan 1998. ISSN 0270-6474.

[157] Martin Svéda, Markéta Castorálová, Jan Lipov, Tomáš Ruml, and Zdeněk Knejzlík. Human UBL5 protein interacts with coilin and meets the Cajal bodies. *Biochemical and biophysical research communications*, 436(2):240–5, jun 2013. ISSN 1090-2104. doi: 10.1016/j.bbrc.2013.05.083.

[158] Juan Wu, Xinhui Liu, Jinjin Fan, Wenfang Chen, Juan Wang, Youjia Zeng, Xiaorang Feng, Xueqing Yu, and Xiao Yang. Bardoxolone methyl (BARD) ameliorates aristolochic acid (AA)-induced acute kidney injury through Nrf2 pathway. *Toxicology*, 318:22–31, apr 2014. ISSN 1879-3185. doi: 10.1016/j.tox.2014.01.008.

[159] Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319, 2002. ISSN 10614036. doi: 10.1038/ng895.

[160] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue): D808–15, jan 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1094.

[161] Nikolas Papanikolaou, Georgios A. Pavlopoulos, Theodosios Theodosiou, and Ioannis Iliopoulos. Protein–protein interaction predictions using text mining methods. *Methods*, 74:47–53, mar 2015. ISSN 10462023. doi: 10.1016/j.ymeth.2014.10.026.

[162] Russ B. Altman. PharmGKB: A logical home for knowledge relating genotype to drug response phenotype [2], 2007. ISSN 10614036.

[163] Kimberly Van Auken, Joshua Jaffery, Juancarlos N Chan, Hans-Michael Müller, Paul W Sternberg, Hans-Michael Muller, and Paul W Sternberg. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) cellular component curation. *BMC bioinformatics*, 10(1):228, jan 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-228.

[164] Alberto Santos, Kalliopi Tsafou, Christian Stolte, Sune Pletscher-Frankild, Seán I. O'Donoghue, and Lars Juhl Jensen. Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*, 3:e1054, jun 2015. ISSN 2167-8359. doi: 10.7717/peerj.1054.

[165] A. S M Ashique Mahmood, Tsung Jung Wu, Raja Mazumder, and K. Vijay-Shanker. DiMeX: A text mining system for mutation-disease association extraction. *PLoS ONE*, 11(4):1–26, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0152725.

[166] Ayush Singhal, Michael Simmons, and Zhiyong Lu. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Computational Biology*, 12(11):1–19, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005017.

[167] Guido Van Rossum and Fred L Drake. Python Tutorial. *History*, 42(4):1–122, 2010. ISSN 0169-118X. doi: 10.1111/j.1094-348X.2008.00203_7.x.

[168] Eric Sayers. A General Introduction to the E-utilities. 2010.

[169] Chih-Hsuan C.-H. Wei, H.-Y. Hung-Yu Kao, and Zhiyong Lu. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(Web Server issue):W518–22, jul 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt441.

[170] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Research International*, 2015:1–7, aug 2015. ISSN 2314-6133. doi: 10.1155/2015/918710.

[171] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. SR4GN: A Species Recognition Software Tool for Gene Normalization. *PLoS ONE*, 7(6):e38460, jun 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0038460.

[172] Chih Hsuan Wei and Hung Yu Kao. Cross-species gene normalization by species inference. *BMC Bioinformatics*, 12(SUPPL. 8), 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-S8-S5.

[173] Chih Hsuan Wei, Robert Leaman, and Zhiyong Lu. SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedical Text. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1385–1391, 2015. ISSN 21682194. doi: 10.1109/JBHI.2015.2422651.

[174] Sunghwan Sohn, Donald C. Comeau, Won Kim, and John W. Wilbur. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 9, 2008. ISSN 14712105. doi: 10.1186/1471-2105-9-402.

[175] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39(Database issue):D507–13, jan 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq968.

[176] Melissa A Haendel, James P Balhoff, Frederic B Bastian, David C Blackburn, Judith A Blake, Yvonne Bradford, Aurelie Comte, Wasila M Dahdul, Thomas A Dececchi, Robert E Druzinsky, Terry F Hayamizu, Nizar Ibrahim, Suzanna E Lewis, Paula M Mabee, Anne Niknejad, Marc Robinson-Rechavi, Paul C Sereno, and Christopher J Mungall. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics*, 5(1):21, may 2014. ISSN 2041-1480. doi: 10.1186/2041-1480-5-21.

[177] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, and Lars Juhl Jensen. DISEASES: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015. ISSN 10462023. doi: 10.1016/j.ymeth.2014.11.020.

[178] Michael H Kagey, Tiffany A Melhuish, and David Wotton. The polycomb protein Pc2 is a SUMO E3. *Cell*, 113(1):127–37, apr 2003. ISSN 0092-8674.

[179] 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, jan 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1237.

[180] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David Mcclosky. The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[181] Django Software Foundation. Django: The Web framework for perfectionists with deadlines. *Djangoproject.Com*, pages 1–3, 2013.

[182] Michael Owens. *The definitive guide to SQLite*. 2010. ISBN 1590596730. doi: 10.1007/978-1-4302-0172-4.

[183] Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H. Blicher, Christian von Mering, Lars J. Jensen, and Peer Bork. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Research*, 42(D1):D401–D407, jan 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1207.

[184] J. Pfaff, J. Hennig, F. Herzog, R. Aebersold, M. Sattler, D. Niessing, and G. Meister. Structural features of Argonaute-GW182 protein interactions. *Proceedings of the National Academy of Sciences*, 110(40):E3770–E3779, oct 2013. ISSN 0027-8424. doi: 10.1073/pnas.1308510110.

[185] Byoung Kwon Yoo, Prasanna K. Santhekadur, Rachel Gredler, Dong Chen, Luni Emdad, Sujit Bhutia, Lewis Pannell, Paul B. Fisher, and Devanand Sarkar. Increased RNA-induced silencing complex (RISC) activity contributes to hepatocellular carcinoma. *Hepatology*, 53(5):1538–1548, 2011. ISSN 02709139. doi: 10.1002/hep.24216.

[186] Gunter Meister. Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*, 14(7):447–459, jul 2013. ISSN 1471-0056. doi: 10.1038/nrg3462.

[187] H. Jin, M. R. Suh, J. Han, K.-H. Yeom, Y. Lee, I. Heo, M. Ha, S. Hyun, and V. N. Kim. Human UPF1 Participates in Small RNA-Induced mRNA Downregulation. *Molecular and Cellular Biology*, 29(21):5789–5799, 2009. ISSN 0270-7306. doi: 10.1128/MCB.00653-09.

[188] Lasse Weinmann, Julia Höck, Tomi Ivacevic, Thomas Ohrt, Jörg Mütze, Petra Schwille, Elisabeth Kremmer, Vladimir Benes, Henning Urlaub, and Gunter Meister. Importin 8 Is a Gene Silencing Factor that Targets Argonaute Proteins to Distinct mRNAs. *Cell*, 136(3):496–507, 2009. ISSN 00928674. doi: 10.1016/j.cell.2008.12.023.

[189] Benjamin Czech and Gregory J. Hannon. Small RNA sorting: Matchmaking for argonautes, 2011. ISSN 14710056.

[190] Justin M Pare, Nasser Tahbaz, Joaquín López-Orozco, Paul LaPointe, Paul Lasko, and Tom C Hobman. Hsp90 regulates the function of argonaute 2 and its recruitment to stress granules and P-bodies. *Molecular biology of the cell*, 20(14):3273–84, jul 2009. ISSN 1939-4586. doi: 10.1091/mbc.E09-01-0082.

[191] Nasser Tahabaz, Fabrice A Kolb, Haidi Zhang, Katarzyma Jaronczyk, Witold Filipowicz, and Tom C Hobman. Characterization of the interactions between mammalian PAZ PIWI domain proteins and Dicer. *EMBO Reports*, 5(2):189–194, feb 2004. ISSN 1469221X. doi: 10.1038/sj.embor.7400070.

[192] Yoontae Lee, Inha Hur, Seong Yeon Park, Young Kook Kim, Ra Suh Mi, and V. Narry Kim. The role of PACT in the RNA silencing pathway. *EMBO Journal*, 25(3): 522–532, 2006. ISSN 02614189. doi: 10.1038/sj.emboj.7600942.

[193] S. S. Truesdell, R. D. Mortensen, M. Seo, J. C. Schroeder, J. H. Lee, O. LeTonqueze, and S. Vasudevan. MicroRNA-mediated mRNA Translation Activation in Quiescent Cells and Oocytes Involves Recruitment of a Nuclear microRNP. *Scientific Reports*, 2 (1):842, dec 2012. ISSN 2045-2322. doi: 10.1038/srep00842.

[194] T. Eystathioy. A Phosphorylated Cytoplasmic Autoantigen, GW182, Associates with a Unique Population of Human mRNAs within Novel Cytoplasmic Speckles. *Molecular Biology of the Cell*, 13(4):1338–1351, 2002. ISSN 10591524. doi: 10.1091/mbc.01-11-0544.

[195] Cátia M. Pereira, Evelyn Sattlegger, Hao Yuan Jiang, Beatriz M. Longo, Carolina B. Jaqueta, Alan G. Hinnebusch, Ronald C. Wek, Luiz E A M Mello, and Beatriz A. Castilho. IMPACT, a protein preferentially expressed in the mouse brain, binds GCN1 and inhibits GCN2 activation. *Journal of Biological Chemistry*, 280(31):28316–28323, 2005. ISSN 00219258. doi: 10.1074/jbc.M408571200.

[196] Johann Bergholz and Zhi Xiong Xiao. Role of p63 in development, tumorigenesis and cancer progression. *Cancer Microenvironment*, 2012. ISSN 18752284. doi: 10.1007/s12307-012-0116-9.

[197] Yijun Zhang, Miaomiao Fan, Xue Zhang, Feng Huang, Kang Wu, Junsong Zhang, Jun Liu, Zhuoqiong Huang, Haihua Luo, Liang Tao, and Hui Zhang. Cellular microRNAs up-regulate transcription via interaction with promoter TATA-box motifs. *RNA*, 20 (12):1878–1889, 2014. ISSN 1355-8382. doi: 10.1261/rna.045633.114.

[198] Jianjun Hu, Fengchao Wang, Xiaoquan Zhu, Ye Yuan, Mingxiao Ding, and Shaorong Gao. Mouse ZAR1-Like (XM-359149) colocalizes with mRNA processing components and its dominant-negative mutant caused two-cell-stage embryonic arrest. *Developmental Dynamics*, 239(2):407–424, 2010. ISSN 10588388. doi: 10.1002/dvdy.22170.

[199] Gunter Meister, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell*, 15(2):185–197, 2004. ISSN 10972765. doi: 10.1016/j.molcel.2004.07.007.

[200] Inga Loedige, Dimos Gaidatzis, Ragna Sack, Gunter Meister, and Witold Filipowicz. The mammalian TRIM-NHL protein TRIM71/LIN-41 is a repressor of mRNA function. *Nucleic Acids Research*, 41(1):518–532, jan 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1032.

[201] P. K. Phalora, N. M. Sherer, S. M. Wolinsky, C. M. Swanson, and M. H. Malim. HIV-1 Replication and APOBEC3 Antiviral Activity Are Not Regulated by P Bodies. *Journal of Virology*, 86(21):11712–11724, 2012. ISSN 0022-538X. doi: 10.1128/JVI.00595-12.

[202] Victoria James, Yining Zhang, Daniel E. Foxler, Cornelia H. de Moor, Yi Wen Kong, Thomas M. Webb, Tim J. Self, Yungfeng Feng, Dimitrios Lagos, Chia-Ying Chu, Tariq M. Rana, Simon J. Morley, Gregory D. Longmore, Martin Bushell, and Tyson V. Sharp. LIM-domain proteins, LIMD1, Ajuba, and WTIP are required for microRNA-mediated gene silencing. *Proceedings of the National Academy of Sciences*, 107(28):12499–12504, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0914987107.

[203] Incheol Ryu, Ji Hoon Park, Sihyeon An, Oh Sung Kwon, and Sung Key Jang. eIF4GI Facilitates the MicroRNA-Mediated Gene Silencing. *PLoS ONE*, 8(2), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0055725.

[204] Marianthi Kiriakidou, Grace S. Tan, Styliani Lamprinaki, Mariangels De Planell-Saguer, Peter T. Nelson, and Zissimos Mourelatos. An mRNA m7G Cap Binding-like Motif within Human Ago2 Represses Translation. *Cell*, 129(6):1141–1151, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.05.016.

[205] Mayuko Yoda, Tomoko Kawamata, Zain Paroo, Xuecheng Ye, Shintaro Iwasaki, Qinghua Liu, and Yukihide Tomari. ATP-dependent human RISC assembly pathways. *Nature Structural and Molecular Biology*, 17(1):17–24, 2010. ISSN 15459985. doi: 10.1038/nsmb.1733.

[206] Yoshiko Maida, Mami Yasukawa, Miho Furuuchi, Timo Lassmann, Richard Possemato, Naoko Okamoto, Vivi Kasim, Yoshihide Hayashizaki, William C. Hahn, and Kenkichi Masutomi. An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature*, 461(7261):230–235, 2009. ISSN 00280836. doi: 10.1038/nature08283.

[207] Ho Young Lee, Kaihong Zhou, Alison Marie Smith, Cameron L. Noland, and Jennifer A. Doudna. Differential roles of human Dicer-binding proteins TRBP and PACT in small RNA processing. *Nucleic Acids Research*, 41(13):6568–6576, 2013. ISSN 03051048. doi: 10.1093/nar/gkt361.

[208] Hong Wei Wang, Cameron Noland, Bunpote Siridechadilok, David W. Taylor, Enbo Ma, Karin Felderer, Jennifer A. Doudna, and Eva Nogales. Structural insights into RNA processing by the human RISC-loading complex. *Nature Structural and Molecular Biology*, 16(11):1148–1153, 2009. ISSN 15459993. doi: 10.1038/nsmb.1673.

[209] I. J. MacRae, E. Ma, M. Zhou, C. V. Robinson, and J. A. Doudna. In vitro reconstitution of the human RISC-loading complex. *Proceedings of the National Academy of Sciences*, 105(2):512–517, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0710869105.

[210] G. Brett Robb and Tariq M. Rana. RNA Helicase A Interacts with RISC in Human Cells and Functions in RISC Loading. *Molecular Cell*, 26(4):523–537, 2007. ISSN 10972765. doi: 10.1016/j.molcel.2007.04.016.

[211] Thimmaiah P. Chendrimada, Richard I. Gregory, Easwari Kumaraswamy, Jessica Norman, Neil Cooch, Kazuko Nishikura, and Ramin Shiekhattar. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436 (7051):740–744, 2005. ISSN 00280836. doi: 10.1038/nature03868.

[212] Motoaki Wakiyama, Koji Takimoto, Osamu Ohara, and Shigeyuki Yokoyama. Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes and Development*, 21(15):1857–1862, 2007. ISSN 08909369. doi: 10.1101/gad.1566707.

[213] A. K. L. Leung, J. M. Calabrese, and P. A. Sharp. Quantitative analysis of Argonaute protein reveals microRNA-dependent localization to stress granules. *Proceedings of the National Academy of Sciences*, 103(48):18125–18130, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0608845103.

[214] E. Prodromaki, A. Korpetinou, E. Giannopoulou, E. Vlotinou, M. Chatziathanasiadou, N. I. Papachristou, C. D. Scopa, H. Papadaki, H. P. Kalofonos, and D. J. Papachristou. Expression of the microRNA regulators Drosha, Dicer and Ago2 in non-small cell lung carcinomas. *Cellular Oncology*, 38(4):307–317, aug 2015. ISSN 2211-3428. doi: 10.1007/s13402-015-0231-y.

[215] A. Detzer, M. Overhoff, A. Mescalchin, M. Rompf, and G. Sczakiel. Phosphorothioate-stimulated cellular uptake of siRNA: a cell culture model for mechanistic studies. *Current pharmaceutical design*, 14(34):3666–73, dec 2008. ISSN 1873-4286. doi: 10.2174/138161208786898770.

[216] Bahnisikha Barman and Suvendra N. Bhattacharyya. mRNA Targeting to Endoplasmic Reticulum Precedes Ago Protein Interaction and MicroRNA (miRNA)-mediated Translation Repression in Mammalian Cells. *Journal of Biological Chemistry*, 290 (41):24650–24656, oct 2015. ISSN 0021-9258. doi: 10.1074/jbc.C115.661868.

[217] D. TEIXEIRA. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA*, 11(4):371–382, 2005. ISSN 1355-8382. doi: 10.1261/rna.7258505.

[218] Arnaud Hubstenberger, Maïté Courel, Marianne Bénard, Sylvie Souquere, Michèle Ernoult-Lange, Racha Chouaib, Zhou Yi, Jean-Baptiste Morlot, Annie Munier, Magali Fradet, Maëlle Daunesse, Edouard Bertrand, Gérard Pierron, Julien Mozziconacci, Michel Kress, and Dominique Weil. P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Molecular Cell*, 68(1):144–157.e5, oct 2017. ISSN 10972765. doi: 10.1016/j.molcel.2017.09.003.

[219] Yun Ju Kim, Alexis Maizel, and Xuemei Chen. Traffic into silence: endomembranes and post-transcriptional RNA silencing. *The EMBO journal*, 33(9):968–80, may 2014. ISSN 1460-2075. doi: 10.1002/embj.201387262.

[220] Andrew J McKenzie, Daisuke Hoshino, Nan Hyung Hong, Diana J Cha, Jeffrey L Franklin, Robert J Coffey, James G Patton, and Alissa M Weaver. KRAS-MEK Signaling Controls Ago2 Sorting into Exosomes. *Cell Reports*, 15(5):978–987, may 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.03.085.

[221] D E Cikaluk, N Tahbaz, L C Hendricks, G E DiMattia, D Hansen, D Pilgrim, and T C Hobman. GERp95, a membrane-associated protein that belongs to a family of proteins involved in stem cell differentiation. *Molecular biology of the cell*, 10(10): 3357–72, oct 1999. ISSN 1059-1524.

[222] Nasser Tahbaz, Jon B. Carmichael, and Tom C. Hobman. GERp95 Belongs to a Family of Signal-transducing Proteins and Requires Hsp90 Activity for Stability and Golgi Localization. *Journal of Biological Chemistry*, 276(46):43294–43299, 2001. ISSN 00219258. doi: 10.1074/jbc.M107808200.

[223] Lukas Stalder, Wolf Heusermann, Lena Sokol, Dominic Trojer, Joel Wirz, Justin Hean, Anja Fritzsche, Florian Aeschimann, Vera Pfanzagl, Pascal Basselet, Jan Weiler, Martin Hintersteiner, David V Morrissey, and Nicole C Meisner-Kober. The rough endoplasmatic reticulum is a central nucleation site of siRNA-mediated RNA silencing. *The EMBO Journal*, 32(8):1115–1127, mar 2013. ISSN 0261-4189. doi: 10.1038/emboj.2013.52.

[224] Joanna J Moser, Marvin J Fritzler, and Jerome B Rattner. Repression of GW/P body components and the RNAi microprocessor impacts primary ciliogenesis in human astrocytes. *BMC cell biology*, 12:37, aug 2011. ISSN 1471-2121. doi: 10.1186/1471-2121-12-37.

[225] Dhimankrishna Ghosh, Dustin Lippert, Oleg Krokhin, John P. Cortens, and John A. Wilkins. Defining the membrane proteome of NK cells. *Journal of Mass Spectrometry*, 45(1):1–25, jan 2010. ISSN 10765174. doi: 10.1002/jms.1696.

[226] Lu Zhu and Martin Ester. Bayesian Collective Markov Random Fields for Subcellular Localization Prediction of Human Proteins. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*, pages 321–329, Boston, MA, USA, 2017. ACM Press. ISBN 9781450347228. doi: http://dx.doi.org/10.1145/3107411.3107412.

[227] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564, 2012. ISSN 1471-0056. doi: 10.1038/nrg3244.

[228] Ronit Ilouz, Varda Lev-Ram, Eric A Bushong, Travis L Stiles, Dinorah Friedmann-Morvinski, Christopher Douglas, Geoffrey Goldberg, Mark H Ellisman, and Susan S Taylor. Isoform-specific subcellular localization and function of protein kinase A identified by mosaic imaging of mouse brain. 2017. doi: 10.7554/eLife.17681.001.

[229] Anna-Lisa Paul, Paul C. Sehnke, and Robert J. Ferl. Isoform-specific Subcellular Localization among 14-3-3 Proteins in Arabidopsis Seems to be Driven by Client Interactions. *Molecular Biology of the Cell*, 16(4):1735, apr 2005. ISSN 1059-1524. doi: 10.1091/MBC.E04-09-0839.

[230] Marija Buljan, Guilhem Chalancon, Sebastian Eustermann, Gunter P Wagner, Monika Fuxreiter, Alex Bateman, and M Madan Babu. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell*, 46(6):871–83, jun 2012. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.05.039.

[231] William J Israelsen and Matthew G Vander Heiden. Pyruvate kinase: Function, regulation and role in cancer. *Seminars in cell and developmental biology*, 43:43–51, jul 2015. ISSN 1096-3634. doi: 10.1016/j.semcdb.2015.08.004.

[232] Masamitsu Konno, Hideshi Ishii, Jun Koseki, Nobuhiro Tanuma, Naohiro Nishida, Koichi Kawamoto, Tatsunori Nishimura, Asuka Nakata, Hidetoshi Matsui, Kozou Noguchi, Miyuki Ozaki, Yuko Noguchi, Hiroshi Shima, Noriko Gotoh, Hiroaki Nagano, Yuichiro Doki, and Masaki Mori. Pyruvate kinase M2, but not M1, allele maintains immature metabolic states of murine embryonic stem cells. *Regenerative Therapy*, 1:63–71, jun 2015. ISSN 2352-3204. doi: 10.1016/J.RETH.2015.01.001.

Notation

Abbreviation

MTDH LYRIC protein

AGO2 argonaute-2

API application programming interface

ATP adenosine triphosphate

BCMRFs Bayesian collective Markov Random Fields

BR binary relevance

BTO BRENDA Tissue Ontology

DDX20 probable ATP-dependent RNA helicase DDX20

DHX15 pre-mRNA-splicing factor ATP-dependent RNA helicase DHX15

DICER1 endoribonuclease dicer protein

ECC edge clustering coefficient

ER endoplasmic reticulum

FP false positive

GCN1 eIF-2-alpha kinase activator GCN1

GCNA gene co-expression network analysis

GEMI4 Gem-associated protein 4

GO Gene Ontology

GO-CCO cellular component ontology

HPA Human Protein Atlas

HS90 heat shock protein 90

IE information extraction

IF immunofluorescence

IPO8 importin 8

JSON JavaScript Object Notation

KLR kernel-based logistic regression

*k***NN** *k* nearest neighbor

LP label powerset

MAP maximum a posteriori

MC mitochondria

MCL Markov clustering

MCMC Monte Carlo Markov Chain

MFP multi-functional protein

miRNA micro RNA

MLC multi-label classification

MLD multi-labeled dataset

MLE maximum likelihood estimation

ML multi-localization

MLP multi-localizing protein

MPLE maximum pseudo-likelihood estimation

MRF Markov random field

mRNA messenger ribonucleic acid

NER mamed entity recognition

NLP natural language processing

PRKRA interferon-inducible double-stranded RNA-dependent protein kinase activator A

PLF pseudo-likelihood function

PMID PubMed-Indexed for MEDLINE

PPI protein-protein interaction

PPIN protein-protein interaction network

RISC RNA-induced silencing complex

RLC RISC loading complex

RNA ribonucleic acid

ROC receiver operating characteristic

rRNA ribosomal ribonucleic acid

SCC subcellular compartment

SCL subcellular localization

siRNA short interfering RNA

SMN complex the survival of motor neurons complex

snRNP small nuclear ribonucleo proteins

SVM support vector machine

TARBP2 RISC-loading complex subunit TARBP2

TP true positive

tissue-specific tissue-specific

Uberon Uber-anatomy ontology

UniProtKB UniProt Knowledgebase

UPF1 protein regulator of nonsense transcripts 1

Nomenclature

α prior probability of protein localized in a SCL

β parameter for potential of interacting proteins

η parameter for potential of protein features

μ parameter for potential of spatial adjacency

ω confidential score of the interaction

ϕ potential

θ parameter set

C count of extracted association

D data set

E energy function

F a vector of protein features

I iteration

i, j protein index

L label set

r random variable follows uniform distribution

R cooling rate

T temperature

w_a, w_s weight for co-occurrence within abstract and sentence

z normalized score

A adjacent matrix

G graph

V vertex