

United Arab Emirates University

College of Information Technology

PROTEIN DOMAIN LINKER PREDICTION: A DIRECTION FOR  
DETECTING PROTEIN-PROTEIN INTERACTIONS

Maad Shatnawi

This dissertation is submitted in partial fulfilment of the requirements for the degree  
of Doctor of Philosophy

Under the Supervision of Dr. Nazar Zaki

June 2015

### **Declaration of Original Work**

I, Maad Shatnawi, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this dissertation entitled “*Protein Domain Linker Prediction: A Direction for Detecting Protein-Protein Interactions*”, hereby, solemnly declare that this dissertation is an original research work that has been done and prepared by me under the supervision of Dr. Nazar Zaki, in the College of Information Technology at UAEU. This work has not been previously formed as the basis for the award of any academic degree, diploma or a similar title at this or any other university. The materials borrowed from other sources and included in my dissertation have been properly cited and acknowledged.

Copyright © 2015 Maad Shatnawi  
All Rights Reserved

## Approval of the Doctorate Dissertation

This Doctorate Dissertation is approved by the following Examining Committee Members:

- 1) Advisor (Committee Chair): Dr. Nazar Zaki

Title: Associate Professor

Department of Intelligent Systems

College of Information Technology

Signature \_\_\_\_\_ Date \_\_\_\_\_

- 2) Member: Dr. Hany Al Ashwal

Title: Assistant Professor

Department of Intelligent Systems

College of Information Technology

Signature \_\_\_\_\_ Date \_\_\_\_\_

- 3) Member: Professor Amr Amin

Title: Professor

Department of Biology

College of Science

Signature \_\_\_\_\_ Date \_\_\_\_\_

- 4) Member (External Examiner): Professor Hesham H. Ali

Title: Professor

Department of Information Science and Technology

Institution: University of Nebraska Omaha (UNOmaha)

Signature \_\_\_\_\_ Date \_\_\_\_\_

This Doctorate Dissertation is accepted by:

Dean of the College of Information Technology: Dr. Shayma Al Kobaisi

Signature \_\_\_\_\_ Date \_\_\_\_\_

Dean of the College of the Graduate Studies: Professor Nagi T. Wakim

Signature \_\_\_\_\_ Date \_\_\_\_\_

Copy \_\_\_\_ of \_\_\_\_

## Abstract

Protein chains are generally long and consist of multiple domains. Domains are the basic elements of protein structures that can exist, evolve, and function independently. The accurate and reliable identification of protein domains and their interactions has very important impacts in several protein research areas. The accurate prediction of protein domains is a fundamental stage in both experimental and computational proteomics. The knowledge of domains is an initial stage of protein tertiary structure prediction which can give insight into the way in which proteins work. The knowledge of domains is also useful in classifying proteins, understanding their structures, functions and evolution, and predicting protein-protein interactions (PPI). However, predicting structural domains within proteins is a challenging task in computational biology. A promising direction of domain prediction is detecting inter-domain linkers and then predicting the regions of the protein sequence in which the structural domains are located accordingly.

Protein-protein interactions occur at almost every level of cell function. The identification of interaction among proteins and their associated domains provide a global picture of cellular functions and biological processes. It is also an essential step in the construction of PPI networks for human and other organisms. PPI prediction has been considered as a promising alternative to the traditional drug design techniques. The identification of possible viral-host protein interactions can lead to a better understanding of infection mechanisms and, in turn, to the development of several medication drugs and treatment optimization.

In this work, a compact and accurate approach for inter-domain linker prediction is developed based solely on protein primary structure information. Then,

inter-domain linker knowledge is used in predicting structural domains and detecting PPI. The research work in this dissertation can be summarized in three main contributions. The first contribution is predicting protein inter-domain linker regions by introducing the concept of amino acid compositional index and refining the prediction by using the Simulated Annealing optimization technique. The second contribution is identifying structural domains based on inter-domain linker knowledge. The inter-domain linker knowledge, represented by the compositional index, is enhanced by the incorporation of biological knowledge, represented by amino acid physiochemical properties, to develop a well-optimized Random Forest classifier for predicting novel domains and inter-domain linkers. In the third contribution, the domain information knowledge is utilized to predict protein-protein interactions. This is achieved by characterizing structural domains within protein sequences, analyzing their interactions, and predicting protein interactions based on their interacting domains. The experimental studies and the higher accuracy achieved is a valid argument in favor of the proposed framework.

**Keywords:** Protein domain identification, domain-linker prediction, compositional index, physiochemical properties, protein-protein interaction prediction, PPI, domain-domain interactions.

## Title and Abstract (in Arabic)

### التنبؤ بمواقع روابط النطاقات البروتينية كأسلوب للكشف عن التفاعلات البينية

#### للبروتينات

#### الملخص

السلاسل البروتينية غالباً ما تكون طويلة وتتكون من عدة نطاقات بنائية ( structural domains)، وتعرف النطاقات بأنها الوحدات الأساسية في بناء البروتينات والتي بإمكانها التطور وأداء المهام باستقلالية، وان تحديد مواقع نطاقات البروتينات وتفاعلاتها البينية ( protein interactions) بطريقة دقيقة وموثوقة ل ذو أهمية بالغة في مجالات عديدة من البحوث المتعلقة بالبروتينات، ويعتبر التنبؤ الدقيق بمواقع روابط النطاقات (domain linkers) خطوة أساسية في كل من الأساليب المعملية والحاسوبية في الدراسات المتعلقة بالبروتينات، ان معرفة نطاقات البروتينات له دور هام في التعرف على الهياكل البنائية لهذه البروتينات وكيفية عملها وتصنيفها والتنبؤ بتفاعلاتها البينية (protein-protein interactions)، غير أن التنبؤ بتحديد مواقع النطاقات يعتبر مهمة صعبة في علم الأحياء الحاسوبي، وان هناك توجه واعد لتحديد مواقع النطاقات من خلال اكتشاف مواقع روابط هذه النطاقات.

وتتفاعل البروتينات فيما بينها على جميع مستويات الوظائف الحيوية للخلايا، وان تحديد التفاعلات البينية للبروتينات والنطاقات المتعلقة بهذه التفاعلات ليوفر صورة شاملة لوظائف الخلايا الحية والعمليات الحيوية، كما أنه يعتبر خطوة أساسية في انشاء شبكات التفاعلات البينية للبروتينات في الانسان وسائر الكائنات الحية، وقد أصبح التنبؤ بالتفاعلات البينية للبروتينات بديلاً واعداً للأساليب التقليدية في صناعة الأدوية، كما أن تحديد التفاعلات بين بروتينات الشخص المريض والفيروسات المسببة للمرض قد يؤدي الى فهم أعمق لآليات العدوى، وبالتالي الى تطوير العقاقير الطبية وتحسين أساليب العلاج.

وقد قمنا في هذا العمل بتطوير طريقة دقيقة وفعالة تجمع ما بين التنبؤ بمواقع النطاقات البنائية وروابطها والتنبؤ بالتفاعلات البينية للبروتينات من خلال معرفة سلاسل الأحماض الأمينية لهذه البروتينات، ويمكن ايجاز هذا العمل في ثلاثة اسهامات رئيسية، ويتمثل الاسهام الأول في التنبؤ بمواقع روابط النطاقات من خلال تقديمنا لمفهوم المؤشر التركيبي للأحماض الأمينية (amino acid compositional index)، ومن ثم تحسين هذا التنبؤ باستخدام أسلوب محاكاة

تقوية المعادن (Simulated Annealing)، ويتمثل الاسهام الثاني في التنبؤ بمواقع النطاقات البنائية بناء على معرفة مواقع روابط هذه النطاقات، فمواقع الروابط ممثلة بالمؤشر التركيبي للأحماض الأمينية يتم تعزيزها بقيمة بيولوجية ممثلة بالخصائص الفيزيوكيميائية (physiochemical properties) للأحماض الأمينية، لبناء مصنف الغابة العشوائية (Random Forest classifier) للتنبؤ بمواقع النطاقات البنائية، ويتمثل الاسهام الثالث في الاستفادة من معرفة النطاقات في التنبؤ بالتفاعلات البينية للبروتينات عن طريق تحليل التفاعلات البينية للنطاقات (domain-domain interactions) المحتواة في هذه البروتينات ، وقد أثبتت الدراسات التجريبية على دقة التنبؤ العالية لهذا الإطار المقترح.

**الكلمات المفتاحية:** تحديد نطاقات البروتين، التنبؤ بروابط النطاقات، المؤشر التركيبي للأحماض الأمينية، الخصائص الفيزيوكيميائية، التنبؤ بالتفاعلات البينية للبروتينات، التفاعلات البينية للنطاقات.

## **Acknowledgements**

This work is dedicated to the soul of my father who encouraged me to be the best I can be, to have high expectations, and to fight hard for what I believe. To his pure soul that I feel it is always with me encouraging me to complete this journey till its end. I would like to thank my beloved mother for her continuous prayers and encouragement.

Thanks are due to my advisor Dr. Nazar Zaki for his guidance, positive comments, and continuous advices throughout this research. Thanks to the advising committee for their valuable discussions during this work. My thanks are extended to all faculty and staff members at the College of Information Technology and the College of Graduate Studies at the United Arab Emirates University. And before all, thanks to Allah for his help and guidance.

## **Dedication**

*To the soul of my Father who encouraged me to be the best I can be,  
To my beloved Mother for her continuous prayers to me.*

## List of Acronyms and Abbreviations

AA	Amino Acid
AAC	Amino Acid Composition
Ac	Accuracy
ANN	Artificial Neural Networks
CI	Compositional Index
DDI	Domain-Domain Interactions
FN	False Negative
FP	False Positive
HMM	Hidden Markov Models
ML	Machine Learning
MSA	Multiple Sequence Alignments
NMR	Nuclear Magnetic Resonance
nr-PDB	non-redundant Protein Data Bank
NW	Needleman-Wunsch
P	Precision
PDB	Protein Data Bank
PPI	Protein-Protein Interaction
PSSM	Position Specific Score Matrix
R	Recall
RF	Random Forest
SA	Simulated Annealing
Sp	Specificity
SVM	Support Vector Machines
TN	True Negative
TP	True Positive

# Table of Contents

<b>Title</b> . . . . .	<b>i</b>
<b>Declaration of Original Work</b> . . . . .	<b>ii</b>
<b>Copyright</b> . . . . .	<b>iii</b>
<b>Approval of the Doctorate Dissertation</b> . . . . .	<b>iv</b>
<b>Abstract</b> . . . . .	<b>vi</b>
<b>Title and Abstract (in Arabic)</b> . . . . .	<b>viii</b>
<b>Acknowledgments</b> . . . . .	<b>x</b>
<b>Dedication</b> . . . . .	<b>xi</b>
<b>List of Acronyms and Abbreviations</b> . . . . .	<b>xii</b>
<b>Table of Contents</b> . . . . .	<b>xiii</b>
<b>List of Tables</b> . . . . .	<b>xvi</b>
<b>List of Figures</b> . . . . .	<b>xviii</b>
<b>Chapter 1: Introduction</b> . . . . .	<b>1</b>
1.1 Overview . . . . .	1
1.2 Dissertation Outline . . . . .	4

1.3	Background . . . . .	5
1.4	Problem Statement and Motivation . . . . .	11
1.5	Research Objectives . . . . .	12
1.6	Technical Challenges . . . . .	14
<b>Chapter 2:</b>	<b>Related Work . . . . .</b>	<b>18</b>
2.1	Inter-Domain Linker Prediction . . . . .	18
2.1.1	Statistical Methods . . . . .	18
2.1.2	Machine Learning Methods . . . . .	20
2.2	Domain Prediction . . . . .	25
2.2.1	Homology-Based Methods . . . . .	25
2.2.2	Machine Learning Methods . . . . .	28
2.3	Protein-Protein Interaction Prediction . . . . .	32
2.3.1	Sequence-Based Approaches . . . . .	33
2.3.2	Structure-Based Approaches . . . . .	44
<b>Chapter 3:</b>	<b>Research Methodology . . . . .</b>	<b>56</b>
3.1	Method Overview . . . . .	56
3.2	Datasets . . . . .	59
3.2.1	Structural Domains and Inter-Domain Linker Prediction . . . . .	59
3.2.2	PPI Prediction . . . . .	60
3.3	Evaluation Measures . . . . .	63
<b>Chapter 4:</b>	<b>CISA: Inter-Domain Linker Prediction Using Compositional Index and Simulated Annealing . . . . .</b>	<b>65</b>
4.1	Compositional Index . . . . .	66

4.2	Detecting the Optimal Set of Threshold Values Using Simulated Annealing . . . . .	69
4.3	Experimental Results and Discussion . . . . .	73
4.3.1	Performance Comparison . . . . .	79
4.3.2	Biological Relevance . . . . .	79
<b>Chapter 5: Random Forest Approach for Domain and Linker Prediction</b>		<b>83</b>
5.1	Feature Extraction . . . . .	83
5.1.1	Hydrophobicity Profile . . . . .	84
5.1.2	Physiochemical Properties . . . . .	86
5.1.3	Protein Sequence Representation . . . . .	87
5.2	Random Forest Model . . . . .	89
5.3	Experimental Results and Discussion . . . . .	92
5.3.1	Performance Comparison . . . . .	95
5.3.2	Biological Relevance . . . . .	97
<b>Chapter 6: PPI Prediction</b> . . . . .		<b>99</b>
6.1	Method . . . . .	99
6.1.1	Pfam Search . . . . .	100
6.1.2	DDI Database Search . . . . .	103
6.2	Experimental Results and Discussion . . . . .	105
<b>Chapter 7: Conclusions and Future Work Directions</b> . . . . .		<b>111</b>
<b>Bibliography</b> . . . . .		<b>116</b>
<b>List of Publications</b> . . . . .		<b>131</b>

# List of Tables

Table 2.1: Domain-linker prediction approaches. . . . .	23
Table 2.2: Domain prediction approaches. . . . .	30
Table 2.3: Statistical Sequence-based PPI prediction approaches. . . . .	37
Table 2.4: Machine-learning sequence-based PPI prediction approaches. . . . .	45
Table 2.5: Structure-based PPI prediction approaches. . . . .	54
Table 3.1: Summary of domain-linker datasets. . . . .	60
Table 3.2: Protein Tools. . . . .	60
Table 3.3: DDI databases. . . . .	62
Table 4.1: CISA performance comparison using Swiss-Prot/DomCut dataset. . . . .	80
Table 5.1: Hydrophobicity index (kcal/mol) of amino acids in a distribution from non-polar to polar at pH=7 [182]. . . . .	84
Table 5.2: Rose hydrophobicity scale. The scale is correlated to the average area of buried AAs in globular proteins [182]. . . . .	85
Table 5.3: SARAH1 hydrophobicity scale. Each AA is assigned a five-bit code in descending order of the binary value of the corresponding code where the right-half is the negative mirror image of the left-half. The 10 most hydrophobic residues are positive, and the 10 least hydrophobic residues are negative [182]. . . . .	85

Table 5.4: Amino acid classification according to their physiochemical properties [185, 186, 187]. . . . .	87
Table 5.5: Prediction measures after removing features that have less information gain using DS-All dataset. . . . .	95
Table 5.6: Recall, precision, and F-measure using Swiss-Prot/DomCut dataset	96

# List of Figures

Figure 1.1: Amino acid structure [1]. . . . .	6
Figure 1.2: Peptide bond formation and hydrolysis [2]. . . . .	7
Figure 1.3: Schematic diagram of an extended polypeptide chain [2]. . . . .	7
Figure 1.4: Protein secondary structures. . . . .	8
Figure 1.5: Primary, secondary, tertiary, and quaternary structures of a protein. (A) The primary structure is the linear sequence of amino acid residues. (B) The secondary structure indicates the local spatial arrangement of polypeptide backbone yielding an extended $\alpha$ -helical or $\beta$ -sheets. (C) The tertiary structure illustrates the three-dimensional conformation. (D) The quaternary structure indicates the assembly of multiple polypeptide chains [1]. . . . .	9
Figure 1.6: Protein-protein interaction (PDB: 1LFD chain A&B) [3]. . . . .	10
Figure 3.1: Method overview. . . . .	57
Figure 4.1: CISA overview. The method consists of two main steps; calculating the compositional index and then apply Simulated Annealing to refine the prediction by detecting the optimal set of threshold values that distinguish linkers from non-linkers. . . . .	66

- Figure 4.2: Comparison between (a) linker index of [4], (b) compositional index of [5], and (c) the modified compositional index profiles for 1au7\_A protein. The horizontal arrow indicates the actual linker region located in the positions from 74 to 109 (according to NCBI). Each AA below a defined threshold is considered to be a linker. Although all three methods predicted the linker in this case, the linker predicted by CISA is more accurate when compared to the actual linker. . . . . 74
- Figure 4.3: Comparison between (a) linker index of [4], (b) compositional index of [5], and (c) the modified compositional index profiles for 1f6f\_C protein. The horizontal arrow indicates the actual linker region (according to NCBI). Each AA below a defined threshold is considered to be a linker. Although all three methods predicted the linker in this case, the linker predicted by CISA is more accurate when compared to the actual linker. Both linker index of [4] and compositional index of [5] have predicted two clear false linkers (regions  $\approx$  1-10 and 44-60, respectively.) 75
- Figure 4.4: Recall, precision, and F1-measure at a window size of 25 and at different chunk sizes (5 to 36) using DomCut/Swiss-Prot dataset. . . . 76
- Figure 4.5: Recall, precision and F1-measure based on DS-All dataset by [6] and [7]. The sliding window sizes  $w$  is set in the range of 5 to 25 AAs. The average value of the sliding window sizes ( $avg$ ) is also included. . . 77
- Figure 4.6: Recall, precision and F1-measure at a window size of 25 and at different chunk sizes based on DS-All dataset. . . . . 78
- Figure 4.7: CISA performance compared to the state-of-the-art predictors based on the DS-All dataset. . . . . 79

- Figure 4.8: Protein 1au7\_A in DS-All dataset which has 146 AA residues containing two domains. (a) The compositional index (CI) profile (blue) and the optimal threshold values returned by the algorithm (red). The two domains and the linker are shown at the bottom as labeled in the DS-All dataset. It is shown how the algorithm returns threshold values higher than CI in the linker region while it returns threshold values lower than CI in domain regions. (b) The 3D structure for this protein showing the two domains. . . . . 81
- Figure 4.9: The CI profile based on the Breast cancer type 1 susceptibility protein is shown in blue and the optimal threshold values achieved by CISA are shown in red. The three domains according to the NCBI's conserved domain database are represented by the green boxes. It is shown that CISA was able to discriminate between linker regions (threshold > CI) and domain regions (threshold < CI). . . . . 82
- Figure 5.1: Representation of protein sequence by AA features and sliding window. Each protein is replaced by its corresponding AA compositional and physiochemical properties. These property values are then averaged over a window that slides along the length of the protein sequence. . . . 88
- Figure 5.2: Random Forest Algorithm . . . . . 89
- Figure 5.3: Recall, precision, F-measure, and AUC of random forest classifier at different averaging window sizes with fifty protein sequences from DS-All dataset. . . . . 93
- Figure 5.4: Recall, precision, F-measure, and AUC of random forest classifier at different averaging window sizes with fifty protein sequences from DomCut dataset. . . . . 93

Figure 5.5: Number of generated trees optimization. Recall, precision, and F-measure at different number of generated trees performed on DS-All dataset. . . . .	94
Figure 5.6: Recall, precision, and F-measure of six currently available domain boundary/linker predictors compared to our approach performed with DS-All dataset. . . . .	96
Figure 5.7: FAS-associated death domain protein - Q13158 (FADD_HUMAN). The protein contains 208 residues and has two domains and a linker according to RCSB-PDB. Our method succeeded in predicting these two domains as indicated by the orange bars. . . . .	97
Figure 5.8: B-lymphocyte antigen CD19 - P15391 (CD19_HUMAN). The protein contains 556 residues and has two domains and a linker according to RCSB-PDB. Our method succeeded in predicting these two domains as indicated by the orange bars. . . . .	98
Figure 5.9: Izumo sperm-egg fusion protein. The protein contains 194 residues and has one domain according to NCBI. Our method succeeded in predicting this domain as indicated by the orange bar. . . . .	98
Figure 6.1: Overview of the PPI prediction process. . . . .	100
Figure 6.2: Accuracy, sensitivity, and specificity of the state-of-the-art PPI predictors compared to our approach. . . . .	106
Figure 6.3: PPI prediction for YCR077C and YDL160C proteins. Two domains are identified in the first protein and two domains are identified in the second protein. 3did retrieves that domain PF09770 from the first protein interacts with domain PF00270 from the second protein. As a result, the model predicts that the two proteins are interacting. . . . .	108

Figure 6.4: PPI prediction for YDR477W and YER027C proteins. Two domains are identified in the first protein and one domain is identified in the second protein. Domain interactions are retrieved in 3did and IDDI. As a result, the model predicts that the two proteins are interacting. . . . . 109

Figure 6.5: PPI prediction for YDR044W and YCR014C proteins. One domain is identified in the first protein and three domains are identified in the second protein. No domain interactions are retrieved in the DDI database. As a result, the model predicts that the two proteins are not interacting. . . . . 110

# Chapter 1: Introduction

In this chapter, I provide an overview of this work in Section 1.1 followed by the outline of the dissertation in Section 1.2. I provide a background on protein structure in Section 1.3, discuss the problem statement and motivation of the overall research in Section 1.4, illustrate our research objectives in Section 1.5, and discuss the technical challenges in Section 1.6.

## 1.1 Overview

Proteins are essential for cells of all living organisms. The primary structure of a protein is the linear sequence of its amino acid (AA) units. Proteins have several essential biological functions including catalysis of metabolic reactions, make up the structure of tissues, nerve transmission, muscle contraction, cell motility, blood clotting, immunologic defenses, working as hormones and regulatory molecules, and transport of vitamins, minerals, oxygen, and fuels [1].

The basic functional units of proteins are protein domains. Several domains are joined together in different combinations forming multi-domain proteins [8, 9]. Each domain in a protein sequence has its own functions and can work with its neighboring domains to perform certain tasks. Therefore, the development of accurate computational method for splitting proteins into structural domains is vital in protein research [10].

Inter-domain linkers tie neighboring domains and support inter-domain communications in multi-domain proteins. They also provide sufficient flexibility to facilitate domain motions and regulate the inter-domain geometry [11]. Predicting inter-domain linkers has a great importance in precise identification of structural domains within a protein. A promising direction of domain prediction, which will be further investigated in this dissertation, is detecting inter-domain linkers and then predicting the location of structural domain accordingly. This domain knowledge can then be used to understand protein structures, functions and evolution, and to predict protein-protein interactions (PPI). The term “linker” and “inter-domain linker” will be used interchangeably in the dissertation.

A protein interacts with other proteins in order to perform certain tasks. Protein-protein interactions (PPI) occur at almost every level of cell functions. The identification of interactions among proteins provides a global picture of cellular functions and biological processes. Since most biological processes involve one or more PPIs, the accurate identification of the set of interacting proteins in an organism is very useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners [12, 13, 14]. Therefore, the development of accurate and reliable methods for identifying PPIs has very important impacts in several protein research areas and pharmaceutical industry.

The interaction between two proteins usually involves a pair of constituent domains, one from each protein. Therefore, understanding protein interactions at the domain level is crucial to discover unrecognized protein-protein interactions and to enhance drug development [15, 16, 17, 18].

In this work, I use the knowledge of structural domains in predicting protein-protein interactions. However, predicting structural domains is a challenging task in

computational biology. A promising direction to predict the location of structural domain is through predicting inter-domain linkers. Therefore, I propose a novel approach for predicting inter-domain linker regions within proteins using only amino acid sequence information. This is achieved by introducing the concept of amino acid (AA) compositional index. The linker knowledge is then used to identify structural domains. Once structural domains are identified within two protein sequences, I can predict whether these two proteins interact or not by analyzing the interacting structural domains that they contain.

## 1.2 Dissertation Outline

This dissertation is structured as follows. In the rest of this chapter, I provide an overview of protein structure in Section 1.3, discuss the problem statement and motivation of the overall research in Section 1.4, illustrate our research objectives in Section 1.5, and discuss the technical challenges in Section 1.6.

Chapter 2 investigates, categorizes, and compares most of the state-of-the-art computational approaches in linker prediction, domain prediction, and PPI prediction. Chapter 3 provides a comprehensive view of our research methodology in addition to the used datasets and evaluation measures.

Chapter 4 discusses our first contribution which is domain-linker prediction using AA compositional index and simulated annealing. Section 4.1 introduces the proposed formula for AA compositional index. Section 4.2 describes the use of simulated annealing algorithm to refine the domain-linker prediction by detecting the optimal threshold values of AA compositional index.

Chapter 5 describes our second contribution which is the development of a Random Forest machine-learning approach for identifying structural domains based on linker knowledge. Chapter 6 describes our third contribution which is about predicting protein-protein interactions by analyzing their interacting domains.

In chapter 7, I summarize this dissertation and comment on possible future work.

## 1.3 Background

Proteins have several essential biological functions in all living organisms including catalysis of metabolic reactions, make up the structure of tissues, nerve transmission, muscle contraction, cell motility, blood clotting, immunologic defenses, working as hormones and regulatory molecules, and transport of vitamins, minerals, oxygen, and fuels [1]. There are four levels of protein structure which play important role in protein functions. These levels are primary, secondary, tertiary, and quaternary structures.

The primary structure of a protein is the linear sequence of its amino acid (AA) units. Although protein chains can become cross-linked, most polypeptides are unbranched polymers, and therefore, their primary structure can be presented by the AA sequence along their main chain or backbone [19].

AAs consist of carbon, hydrogen, oxygen, and nitrogen atoms that are clustered into functional groups. Each amino acid has a central carbon atom, called the alpha ( $\alpha$ )-carbon, where four different groups are attached to it as shown in Figure 1.1. These groups are the amino group ( $NH_2$ ) and the carboxyl group ( $COOH$ ), a hydrogen atom ( $H$ ), and a distinctive side chain ( $R$ )-group. All amino acids have the same general structure, but each has a different  $R$ -group. The side chains ( $R$ ) are the major determinants of the structure and properties of the AA. The physiochemical characteristics of the amino-acid side chains have important role in the folding and functions of proteins [2].

There are over three hundred naturally occurring AAs on earth, but the number of different AAs in proteins is only twenty. These twenty amino acids are Alanine, Arginine, Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Proline, Serine, Threo-

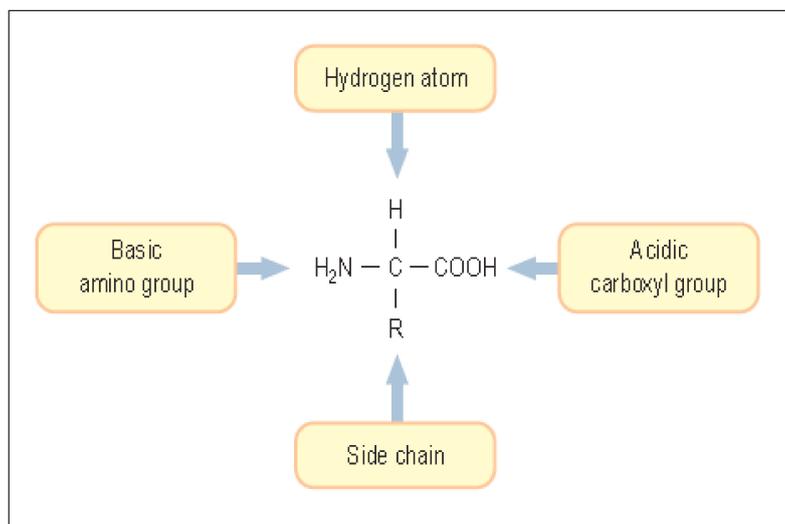


Figure 1.1: Amino acid structure [1].

nine, Tryptophan, Tyrosine, and Valine represented by one-letter abbreviation as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V, respectively.

Amino acids are connected to make proteins by a chemical reaction in which a molecule of water is removed, leaving two amino acids residues connected by a peptide bond [19] as shown in Figure 1.2. Connecting multiple AAs in this way produces a polypeptide as shown in Figure 1.3. This reaction leaves the *C* of the carboxyl group directly linked to the *N* of the amino group. The starting end of the protein with a free amino group is known as the amino terminal (*N*-terminal) whereas the ending end with a free carboxyl group is known as the carboxyl terminal (*C*-terminal). Polypeptides can be thought of as a string of alpha carbons alternating with peptide bonds. Since each alpha carbon is attached to an *R*-group, a given polypeptide is distinguished by the sequence of its *R*-groups.

The secondary structure of a protein is the general three-dimensional form of its local parts. The most common secondary structures are alpha ( $\alpha$ ) helices and beta ( $\beta$ ) sheets. The  $\alpha$ -helix is a right-handed spiral array while the  $\beta$  sheet is made up of beta strands connected crosswise by two or more hydrogen bonds, forming a twisted pleated sheet. These secondary structures are linked together by tight turns and loose flexible

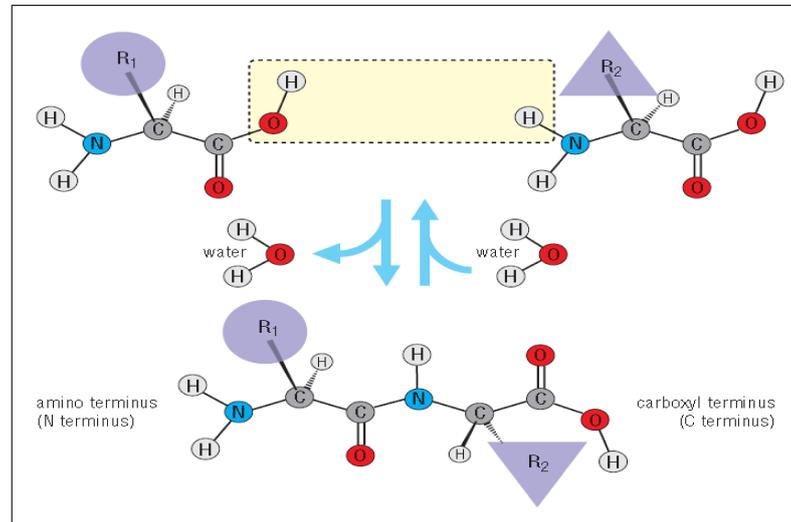


Figure 1.2: Peptide bond formation and hydrolysis [2].

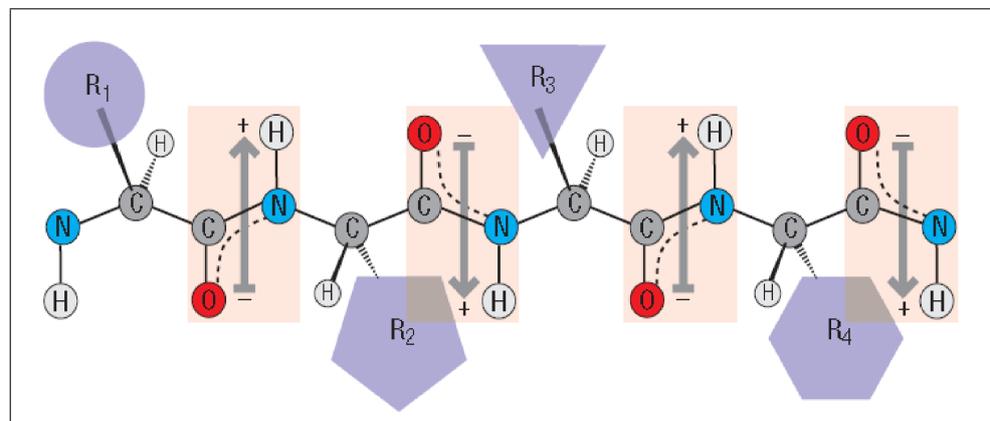


Figure 1.3: Schematic diagram of an extended polypeptide chain [2].

loops [20] as shown in Figure 1.4.

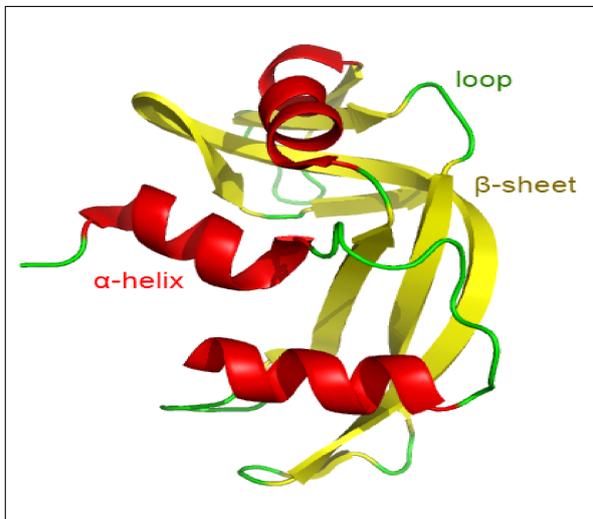


Figure 1.4: Protein secondary structures.  
(<https://www.ocf.berkeley.edu/asiegel/posts/?author=1&paged=4>)

The tertiary structure of a protein is its three-dimensional folded and biologically active conformation which reflects the overall shape of the molecules. The tertiary structure of proteins is determined by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [1]. Domains are the basic functional units of protein tertiary structures. A protein domain is a conserved part of a protein that can evolve, function, and exist independently.

Quaternary structure refers to a complex or an assembly of two or more separate peptide chains that are held together by non-covalent or, in some cases, covalent interactions. Most proteins consist of more than one chain and are referred to as dimeric, trimeric, or multimeric proteins [1]. Figure 1.5 illustrates the four levels of protein structure.

Although many proteins are composed of a single structural domain, most proteins are built up from two or more domains joined together in different combinations [8, 9]. Each domain in a multi-domain protein has its own functions and can work with its neighboring domains to perform certain tasks. One domain may exist in a variety

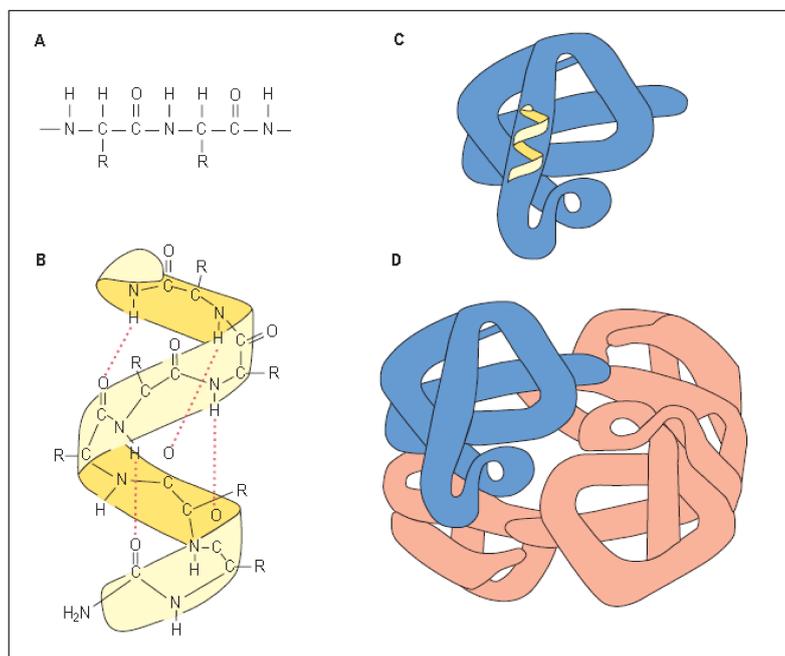


Figure 1.5: Primary, secondary, tertiary, and quaternary structures of a protein. (A) The primary structure is the linear sequence of amino acid residues. (B) The secondary structure indicates the local spatial arrangement of polypeptide backbone yielding an extended  $\alpha$ -helical or  $\beta$ -sheets. (C) The tertiary structure illustrates the three-dimensional conformation. (D) The quaternary structure indicates the assembly of multiple polypeptide chains [1].

of different proteins. The function of the entire protein is determined by the properties of its domains. Domains vary in length from 25 to 500 amino acids [4]. Inter-domain linkers tie neighboring domains and support inter-domain communications in multi-domain proteins. They also provide sufficient flexibility to facilitate domain motions and regulate the inter-domain geometry [11].

Predicting protein functions through protein structure is a complex task. As a result, several methods have recently been developed to predict protein functions using PPI. PPI refers to intentional physical contacts established between two or more proteins through biochemical events and/or electrostatic forces. A protein interacts with other proteins, as illustrated in Figure 1.6, in order to perform certain tasks. PPIs occur at almost every level of cell functions. Most biological processes involve one or more PPIs. Most protein sequences contain multi-domains and the interaction between

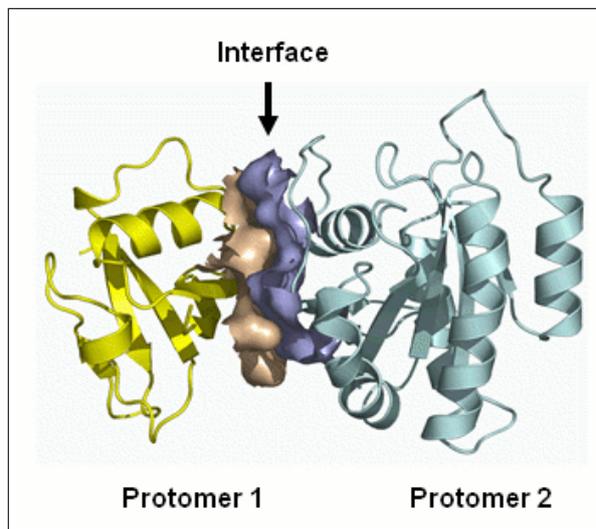


Figure 1.6: Protein-protein interaction (PDB: 1LFD chain A&B) [3].

two proteins usually involves a pair of constituent domains, one from each protein.

## 1.4 Problem Statement and Motivation

The development of an accurate and reliable method for identifying protein domains and their interactions has very important impacts in several protein research areas. The knowledge of domains is an initial stage of protein tertiary structure prediction which can give insight into the way in which proteins work. The knowledge of domains is also useful in classifying proteins, understanding their structures, functions and evolution, and predicting PPIs. However, predicting structural domains is a challenging task in computational biology. A promising direction to predict the location of structural domain is through the prediction of the of the inter-domain linkers. Therefore, the accurate prediction of protein inter-domain linkers is an initial stage in both experimental and computational proteomics.

Since most biological processes involve one or more PPIs, the accurate identification of the set of interacting proteins in an organism is very useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners [14, 12, 13]. Protein interaction prediction is also a fundamental step in the construction of PPI networks for human and other organisms. PPI prediction has been considered as a promising alternative to the traditional drug design techniques. The identification of possible viral-host protein interactions can lead to a better understanding of infection mechanisms and, in turn, to the development of several medication drugs and treatment optimization. In addition, Abnormal PPIs have implications in several neurological disorders such as Creutzfeld-Jacob and Alzheimer [21, 22, 23].

## 1.5 Research Objectives

In this work, a novel and simple method is proposed for predicting inter-domain linker regions within proteins. This is achieved by introducing the concept of AA compositional index. The compositional index is deduced from the protein sequence dataset of domains and linker segments. The compositional index is then enhanced by combining biological knowledge and amino acid physiochemical properties to construct a machine learning-based classifier for predicting novel structural domains and inter-domain linkers. Once structural domains are identified within two protein sequences, it can be predicted whether these two proteins interact or not by analyzing the interacting structural domains they contain.

The main research objectives of this work can be summarized as follows:

- Developing a novel method for identifying domains and inter-domain linkers within protein sequences. This is achieved through the following steps:
  - (1) Predicting protein inter-domain linker regions by utilizing the concept of AA compositional index and refining the prediction using an optimization technique namely Simulated Annealing.
  - (2) Identifying structural domains based on linker knowledge. The linker knowledge, represented by the compositional index, is enhanced by injecting biological knowledge, represented by AA physiochemical properties, to construct a novel protein profile. The protein profile is then used to train a Random Forest classifier for predicting novel domains and inter-domain linkers.
- Developing a PPI prediction method through the following steps:
  - (1) Characterizing domains within protein sequences.

- (2) Identifying interacting domains.
- (3) Predicting protein interactions based on their interacting domains.

## 1.6 Technical Challenges

The proposed method in this dissertation allows a biologist to gain knowledge related to inter-domain linkers, structural domain and eventually the PPI solely from the protein sequence. However, there are several challenges arise from the protein sequence itself. First, there have been a huge amount of newly discovered protein sequences in the post genomic era. Second, protein chains are typically large and contain multiple domains which are difficult to characterize by experimental methods. Third, the availability of large, comprehensive, and accurate benchmark datasets is required for the training and evaluation of prediction methods. Fourth, computational methods are based on experimentally collected data, and therefore, any error in the experimental data will affect the computational predictions.

One of the challenges of prediction methods is the protein representation. The most and simplest model of a protein is its entire amino acid sequence. However, this approach doesn't work well when the query protein does not have high sequence similarity to any known protein [24]. Several statistical-based models were proposed. The simplest statistical model is based on the protein AA composition which is the normalized occurrence frequencies of the twenty amino acids in a protein. However, all the sequence-order knowledge will be lost using this representation which, in turn, will negatively affect the prediction accuracy [24]. Some approaches use amino acid flexibility such as CHOPnet [25], gene ontology, solvent accessibility information, and/or evolutionary information such as DOMpro [26]. Protein secondary structure information has also been broadly used in several domain-linker prediction such as SSEP-Domain [27] and PPI prediction approaches such as PrePPI [28]. However, extracting accurate secondary structure information by itself is another challenge. Protein secondary

structures are normally predicted by SSpro [29] which is an 80% accurate tool, so the incorrectly predicted secondary structures may lead to model misclassification. Many protein prediction approaches such as DomNet [9], PPRODO [30], and DROP [31] uses the Position Specific Score Matrix (PSSM) which requires a high computational cost to be generated. Several approaches have used the 3-D coordinates of protein structure [27].

There are various challenges that face machine-learning protein prediction methods. Selecting the best machine learning approach is a great challenge. There is a variety of techniques that diverse in accuracy, robustness, complexity, computational cost, data diversity, over-fitting, and dealing with missing attributes and different features. Most machine-learning approaches of protein sequence prediction are computationally expensive and often lack high prediction accuracy. They are further susceptible to overfitting. In other words, after a certain point, adding new features or new training examples can reduce the prediction quality [32]. Furthermore, protein chain data are imbalanced as domain regions are much longer than linker regions, and non-interacting protein pairs are much more than interacting pairs, and therefore, classifiers will usually be biased towards the majority class. This raises the challenge of choosing the appropriate evaluation metrics. For example, a technique that fails to predict any linker in a protein sequence which has respectively 95% and 5% of its amino acids as domains and linkers, achieves a high prediction accuracy of as much as 95%. In addition, since highly imbalanced distributions usually lead to large datasets, more efficient prediction methods, algorithmic optimizations and continued improvements in hardware performance are required to handle such challenging tasks.

Some issues for possible further improvements includes capturing long-term AA dependencies and developing a more suitable representation of protein sequence profiles

that includes evolutionary information. Most of the existing approaches showed a limited capability in exploiting long-range interactions that exist among amino acids and participate in the formation of protein secondary and tertiary structure. Residues can be adjacent in 3D space while located far apart in the AA sequence. [9, 33].

One reason behind the limited capability of multi-domain protein predictors is the disagreement of domain assignment within different protein databases. The agreement between domain databases covers about 80% of single domain proteins and only about 66% of multi-domain proteins [34]. This disagreement is due to the variance in the experimental methods used in domain assignment. The most predominant techniques used to experimentally determine protein 3D structures are X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). To determine the conformation of a protein with X-rays, the protein must be in the form of a crystal with a strictly ordered structure. The crystallized protein is then irradiated with X-rays. Protein crystallization is the slowest and most challenging stage in X-ray structural analysis. Some proteins are relatively easy to crystallize within few days, others can take several months or even years, while many proteins such as cell membranes proteins still cannot be crystallized [35]. On the other hand, NMR is based on the fact that some atomic nuclei, such as hydrogen, are intrinsically magnetic. In a magnetic field, these magnetic nuclei can adopt states of different energy. Applying radio-frequency radiation can induce the nuclei to flip between these energy states, which can be measured and depicted in the form of a spectrum [36]. X-ray diffraction has no size limitations and provides more precise atomic detail while information about the dynamics of the molecule may be limited. NMR is the best when no protein crystals can be obtained but it produces lower resolution structures and is generally limited to small molecular weights [37]. This variance in experimental methods of domain assignment can establish

an upper limit for domain-linker prediction accuracy.

## Chapter 2: Related Work

This chapter investigates, classifies, and compares most of the state-of-the-art computational approaches in domain and linker prediction and PPI prediction. Inter-domain linker prediction approaches are discussed in Section 2.1, structural domain prediction approaches are discussed in Section 2.2, and PPI prediction approaches are discussed in Section 2.3.

### 2.1 Inter-Domain Linker Prediction

Several impressive protein inter-domain linker and domain boundary prediction methods have been developed and can be classified into statistical-based and Machine-Learning (ML)-based methods.

#### 2.1.1 Statistical Methods

Statistical-based methods use statistical features of proteins such AA frequencies and AA composition to predict domain-linker regions. Examples of these methods are DomCut [4] and GlobPlot [38].

**DomCut:**

DomCut<sup>1</sup> [4] is one of the typical early day’s statistical-based methods. Domcut predicts domain linker regions based on the differences in AA composition between domain and linker regions in a protein sequence. In their research, a region or segment in a sequence is considered as linker if it is in the range from 10 to 100 residues, connecting two adjacent domains, and not containing membrane spanning regions. To represent the preference for AA residues in linker regions, they defined the linker index as the ratio of the frequency of AA residue in domain regions to that in linker regions:

$$L_i = -\ln\left(\frac{f_i^{linker}}{f_i^{nonlinker}}\right) \quad (2.1)$$

where  $f_i^{linker}$  and  $f_i^{non-linker}$  are the frequencies of amino acid residue  $i$  in linker and non-linker regions, respectively.

A linker preference profile was generated by plotting the averaged linker index values along an AA sequence using a sliding window of size 15 AAs. A linker was predicted if there was a trough in the linker region and the averaged linker index value at the minimum of the trough was lower than the threshold value. At the threshold value of 0.09, the sensitivity and selectivity of DomCut were 53.5% and 50.1%, respectively. Despite the fact that DomCut showed glimpse of potential success, it was reported by Dong *et al.* [39] that DomCut has low sensitivity and specificity in comparison to other recent methods. However, integrating more biological evidences with the linker index could enhance the prediction and therefore, the idea of DomCut was later utilized by several researchers such as Zaki *et al.* [5] and Pang *et al.* [40].

---

<sup>1</sup><http://www.bork.embl.de/suyama/domcut/>

**GlobPlot:**

Linding *et al.* [38] proposed another statistical method called GlobPlot<sup>2</sup> based on protein secondary structure information. GlobPlot allows users to plot the tendency within protein sequences for exploring both potential globular and disordered/flexible regions in proteins based on their AA sequence, and to identify inter-domain segments containing linear motifs.

Other statistical-based methods are Udvary *et al.* [41] which predicts the locations of linker regions within large multi-functional proteins and Armadillo [42] which predicts domain linkers by using AA composition.

### 2.1.2 Machine Learning Methods

Machine learning (ML) based methods are the most commonly used approaches in inter-domain linker prediction. Most of the recent ML approaches employ either Artificial Neural Networks (ANN) or Support Vector Machines (SVM). ANN approaches include PPRODO [30], DomNet [9], and Shandy [43]. SVM approaches include DoBo [44], and DROP [31].

**PPRODO:**

Sim *et al.* [30] introduced PPRODO as an ANN classifier that was trained using features obtained from the Position Specific Scoring Matrix (PSSM) generated by PSI-BLAST. The training dataset contained 522 contiguous two-domain proteins was obtained from the structural classification of proteins (SCOP) database, release version 1.63 [45]. When tested on 48 newly added non-homologous proteins in SCOP version 1.65 and on CASP5 targets, PPRODO achieved 65.5% of prediction accuracy.

---

<sup>2</sup>[http:// globplot.embl.de](http://globplot.embl.de)

One of the limitations of this methods is the high computational cost to generate PSSM.

### **DomNet:**

Yoo *et al.* [9] introduced DomNet (Protein Domain Boundary Prediction Using Enhanced General Regression Network and New Profiles) which was trained using a compact domain profile, secondary structure, solvent accessibility information, inter-domain linker index, evolutionary information, and PSSM to identify possible domain boundaries for a target sequence. The authors proposed a semi-parametric model that uses a nonlinear auto-associative Enhanced General Regression Neural network (EGRN) for filtering noise and less discriminative features. The performance of DomNet was evaluated on the Benchmark2 and CASP7<sup>3</sup> datasets in terms of accuracy, sensitivity, specificity, and correlation coefficient. DomNet achieved an accuracy of 71% for domain boundary determination in multi-domains proteins using Benchmark2 dataset.

One of the advantages of this approach is that EGRN addresses the drawbacks of the General Regression Neural network (GRNN) [46] technique. GRNN is a non-parametric model that requires extensive computer resources by performing very large computations and it suffers from overfitting and burden of dimensionality.

On the other hand, although using structural information could achieve good prediction results, finding the structural information by itself is another challenge. The method requires the computational cost to generate PSSM and to predict secondary structure information for each protein.

### **DROP:**

Ebina *et al.* [31] developed Domain linker pRediction using OPTimal features

---

<sup>3</sup><http://predictioncenter.org/casp7>

(DROP) using a SVM, with an Radial Basis Function (RBF) kernel, inter-domain linker predictor trained by 25 optimal features. The optimal combination of features was selected from a set of 3000 features using a random forest algorithm, which calculates the Mean Decrease Gini Index (MDGI), complemented with a stepwise feature selection. The selected features were primarily related to secondary structures, PSSM elements of hydrophilic residues and prolines.

For each residue, a 3000-dimensional real-valued feature vector was extracted. These features are as follows. 544 AA indices describing physicochemical properties, 20 PSSM elements, three Probabilities of Secondary Structure (PSS), two  $\alpha$ -helix/ $\beta$ -sheet core propensities, one sequential hydrophobic cluster index, sequence complexity as defined by Shannons entropy, one expected contact order, 20 elements of AA compositions, three domain/coil/linker propensity indices, two linker likelihood scores and three newly defined scores quantifying the AA composition similarity between domain and linker regions. Vector elements were averaged with windows of 5, 10, 15 or 20 residues around the considered residue to include local and semi-local information into the vectors. The total number of vectors for linkers and domains were 2230 and 52335, respectively.

The accuracy of DROP was evaluated by two domain linker datasets; DS-All [6, 7], and CASP8 FM<sup>4</sup>. DS-All contains 169 protein sequences, with a maximum sequence identity of 28.6%, and 201 linkers. DROP achieved a prediction sensitivity and precision of 41.3% and 49.4%, respectively, with more than 19.9% improvement by the optimal features. DROP does not use sequence similarity to domain databases. One of the advantages of this approach is the use of random forest approach for feature selection. Instead of exhaustively searching all feature combination, random forest is

---

<sup>4</sup><http://predictioncenter.org/casp8/>

<b>Approach</b>	<b>Extracted Features</b>	<b>Technique/Tool</b>	<b>Datasets</b>
DomCut (Suyama and Ohara 2003)	AA composition	Linker index	Swiss-Prot
GlobPlot (Linding <i>et al.</i> 2003)	Secondary structures	AA propensity	SCOP 1.59
PPRODO (Sim <i>et al.</i> 2005)	PSSM	ANN PSI-BLAST	SCOP 1.65 CASP5
DomNet (Yoo <i>et al.</i> 2008)	Secondary structures, solvent accessibility, linker index, PSSM	EGRN	Benchmark_2 CASP7
DROP (Ebina <i>et al.</i> 2011)	Secondary structures, PSSM	Random Forest, SVM	SCOP 1.65 CASP5

Table 2.1: Domain-linker prediction approaches.

based on random sampling which provides a quick and inexpensive screening for the optimal features. However, DROP overpredicts domain linkers in single-domain targets of Benchmarking DataSet (BDS) [7] and CAFASP4<sup>5</sup>. This can be decreased by increasing the default threshold level or by including non-local features such foldability index. In addition to that, the method requires the computational cost to generate PSSM and to predict secondary structure information for each protein.

Table 2.1 summarize the above mentioned prediction approaches and compares them. Most of the discussed methods have, in general, the following limitations:

- Although methods that use structural information could achieve good prediction results, finding the structural information by itself is another challenge.
- Most of the mentioned methods are computationally expensive as they require the computational cost to generate PSSM and/or predict secondary structure information for each protein.
- Some methods are evaluated based on the overall prediction accuracy only. This may not effectively reflect the issues of the unbalancing problem of protein domain linker data.

---

<sup>5</sup><http://www.cs.bgu.ac.il/~dfischer/CAFASP4/>

In the first contribution of this work, I develop an effective method for inter-domain linker prediction solely from AA sequence information. Domain-linker regions are determined using AA compositional index and then a simulated annealing algorithm is employed to enhance the prediction by finding the optimal threshold value that separates domains from linkers.

## 2.2 Domain Prediction

Structural domain prediction methods can be classified into homology-based, and ML-based methods.

### 2.2.1 Homology-Based Methods

Homology-based methods search the target sequences through known protein structure libraries using alignment, Hidden Markov Models (HMM), or PSI-BLAST techniques. Examples of homology-based methods are CHOP [25], Scooby-Domain [47], DOMpro [26], and FIEFDOM [48], and PFam [49]. Although homology-based methods can achieve high prediction accuracy specially when close templates are retrieved, the accuracy often decreases piercingly when the sequence identity of the target and template is low [50].

#### **DOMpro:**

DOMpro [26] is a typical alignment/homology-based method which requires the use of PSI-BLAST [51] to generate evolutionary and homology information in the form of profiles. DOMpro was independently evaluated along with 12 other predictors in the Critical Assessment of Fully Automated Structure Prediction 4 (CAFASP-4) [52, 53] where it was ranked among the top *ab initio* domain predictors.

#### **Scooby-Domain:**

SequenCe hydrOphOBicitY predicts DOMAINS (Scooby-Domain) web application was developed by George *et al.* [47] and extended by Pang *et al.* [40] to visually identify foldable regions in a protein sequence. Scooby-Domain uses the distribution

of observed lengths and hydrophobicities in domains with known 3D structure to predict novel domains and their boundaries in a protein sequence. It utilizes a multilevel smoothing window to determine the percentage of hydrophobic AAs within a putative domain-sized region in a sequence. Each smoothing window calculates the fraction of hydrophobic residues it encapsulates along a sequence, and places the value at its central position. This creates a triangular-shape 2D matrix where the value at cell  $(i, j)$  is the average hydrophobicity encapsulated by a window of size  $j$  that is centered at residue position  $i$ . Matrix values are converted to probability scores by referring to the observed distribution of domain sizes and hydrophobicities. Using the observed distribution of domain lengths and percentage hydrophobicities, the probability that the region can fold into a domain or be unfolded is then calculated.

Scooby-Domain employs an A\* search algorithm to search through a large number of alternative domain annotations. The A\* search algorithm considers combinations of different domain sizes, using a heuristic function to conduct the search. The corresponding sequence stretch for the first predicted domain is removed from the sequence. The search process is repeated until there are less than 34 residues remaining, which is the size of the smallest domain; or until there are no probabilities greater than 0.33, which is an arbitrary cutoff, to prevent non-domain regions from being predicted as a domain.

Two linker prediction scoring systems, Domcut [4] and PDLI [39], were used separately to complement Scooby-Domains prediction. The performance of Scooby-Domain was evaluated with the inclusion of homology information. Homologues of the query sequence were detected using PSI-BLAST [51] searches of the SWISS-PROT database [54] and *Multiple Sequence Alignments* (MSA) were generated using PRALINE [55]. On a test set of 173 proteins with consensus CATH [56] and SCOP [45] domain

definitions, Scooby-Domain has a sensitivity of 50% and an accuracy of 29%.

The advantages of Scooby-Domain include its ability to predict discontinuous domains and successful predictions are not limited by the length of the query sequence. A\* search is a very flexible method, and it may be easily adapted and improved to include more sophistication in its predictions. . However, A\* search algorithm has an exponential computational time complexity in its worst case [57, 58]. Furthermore, domains that are connected by small linkers may not be identifiable by Scooby-Domain because window averaging may lose any signal at the linker.

#### **FIEDom:**

Bondugula *et al.* [48] presented Fuzzy Integration of Extracted Fragments for Domains (FIEFDom) as a method to predict domain boundaries of a multi-domain protein from its AA sequence using a Fuzzy Mean Operator (FMO). Using the non-redundant (nr) sequence database together with a reference protein set (RPS) containing known domain boundaries, the operator is used to assign a likelihood value for each residue of the query sequence as belonging to a domain boundary. FMO represents a special case of the fuzzy nearest neighbor algorithm [59] with the number of classes set to one. The approach is a three-step procedure. First, the PSSM of the query sequence is generated using a large database of known sequences. Second, the generated profile is used to search for similar fragments in the RPS. Third, the matches with the proteins in RPS are parsed, and the domain Boundary Propensity (PB) of the query protein is predicted using a FMO. For SCOP 1.65 dataset with a maximal sequence identity of 30%, the average domain prediction accuracy of FIEFDom is 97% for one domain proteins and 58% for multi-domain proteins.

The advantages of FMO include its simplicity, ease of updating, and its asymp-

otic error bounds. The choice of the program to designate a region as a domain boundary can be traced back to all proteins in the local database that contributed to the decision. The model doesn't need to be trained or tuned whenever new examples of domain boundaries become available. In addition, the users can choose the domain definitions such as CATH [56] and SCOP [45], to suit their needs by replacing the Reference Protein Set (RPS). FIEFDom works well for protein sequences with many close homologs and that with only remote homologs. On the other hand, this approach did not address the issue of predicting domains with non-contiguous sequences and therefore it discarded such proteins.

### **ThreeDom:**

Xue *et al.* [50] introduced ThreeDom based on multiple threading alignments using a domain conservation score that combines information from template domain structures and terminal and internal alignment gaps. The threading of the target sequences for structural template identifications through the Protein Data Bank (PDB) is performed by LOMETS [60] which is a local meta-threading-server for protein structure prediction.

Although homology-based methods can achieve high prediction accuracy specially when close templates are retrieved, the accuracy often decreases piercingly when the sequence identity of the target and template is low.

## **2.2.2 Machine Learning Methods**

Beside the homology-based methods, there are several ML-based methods for predicting structural domains within proteins. Chatterjee *et al.* [61] and Li *et al.* [62] are examples of such ML-based methods.

**Chatterjee *et al.*:**

Chatterjee *et al.* [61] employed a SVM classifier with three kernel functions; linear, cubic polynomial, and RBF. The feature set consists of six different features; predicted secondary structure, predicted solvent accessibility, predicted conformational flexibility profile, AA composition, PSSM, and AA physicochemical properties. A window of 13 AA long is slid over the protein chain every time by one AA position. The accuracy of this approach was evaluated on CATH datasets [56]. The SVM classifier with a cubic polynomial kernel had shown the best performances in terms of accuracy and precision. These two measures were 76.46% and 86.82% respectively.

**Li *et al.*:**

Li *et al.* [62] proposed a domain prediction method based on combining the techniques of Random Forest, mRMR (maximum relevance minimum redundancy), and IFS (incremental feature selection) and incorporating the features of physicochemical and biochemical properties, sequence conservation, residual disorder, secondary structure, and solvent accessibility. The performance of this approach was evaluated on UniProt/Swiss-Prot database (version 2010.06) [63] and achieved 64.3% sensitivity and 80.8% specificity.

Although using structural information could achieve good prediction results, finding the structural information by itself is another challenge. The above mentioned methods require the computational cost to generate PSSM and to predict secondary structure information for each protein.

Table 2.2 summarize the above mentioned prediction approaches and compares them. Most of the discussed methods have, in general, the following limitations:

Approach	Extracted Features	Technique/Tool	Datasets
DOMpro (Cheng <i>et al.</i> 2006)	Evolutionary and homology information	PSI-BLAST	CAFASP-4
Scooby-Domain (George <i>et al.</i> 2005, Pang <i>et al.</i> 2008)	Domain lengths and hydrophobicities	A*-search	Swiss-Prot
FIEFDom (Bondugula <i>et al.</i> 2009)	PSSM	FMO	SCOP 1.65
ThreeDom (Xue <i>et al.</i> 2013)	Template domain structures, terminal and internal alignment gaps	Multiple threading alignments	CASP8 CASP9 CASP10
Chatterjee <i>et al.</i> (2009)	Secondary structures, solvent accessibility, PSSM, AA composition and physiochemical properties	SVM	CATH
Li <i>et al.</i> (2012)	physicochemical and biochemical properties, sequence conservation residual disorder, secondary structure, solvent accessibility	Random Forest, mRMR , IFS	UniProt/ Swiss-Prot

Table 2.2: Domain prediction approaches.

- Although many ML-based domain predictors have been developed and shown good prediction performance in single-domain proteins, they have shown limited capability in multi-domain proteins [9].
- Although homology-based methods can achieve high prediction accuracy specially when close templates are retrieved, the accuracy often decreases piercingly when the sequence identity of the target and template is low [50].
- Although methods that use structural information could achieve good prediction results, finding the structural information by itself is another challenge.
- Some methods are computationally expensive as they require the computational cost to generate PSSM and/or predict secondary structure information for each protein.

In the second contribution of this work, I develop a simple and effective approach

for predicting structural domains using inter-domain linker knowledge. Inter-domain linkers are generally shorter than domains and can be recognized more simply and efficiently. Recognizing a linker can then lead to discovering two adjacent domains.

## 2.3 Protein-Protein Interaction Prediction

PPI prediction has been studied extensively by several researchers and a large number of approaches have been proposed. These approaches can be classified into physiochemical experimental and computational approaches. Physiochemical experimental techniques identify the physiochemical interactions between proteins which, in turn, are used to predict the functional relationships between them. These techniques include yeast two-hybrid based methods [64], mass spectrometry [65], Tandem Affinity Purification [66], protein chips [67], and hybrid approaches [68]. Although these techniques have succeeded in identifying several important interacting proteins in several species such as Yeast, *Drosophila*, and *Helicobacter-pylori* [69], they are computationally expensive and significantly time consuming, and so far the predicted PPIs have covered only a small portion of the complete PPI network. As a result, the need for computational tools has been increased in order to validate physiochemical experimental results and to predict non-discovered PPIs [14, 70].

Several computational methods have been proposed for PPI prediction and can be classified according to the used protein features into sequence-based and structure-based methods. Sequence-based methods utilize AA features and can be further categorized into statistical and Machine Learning (ML)-based methods. The structure-based methods use three-dimensional structural features [71] and can be categorized into template-based, statistical and ML-based methods. This section provides an overview and discussion of some of the current computational sequence-based and structure-based PPI prediction approaches.

### 2.3.1 Sequence-Based Approaches

Sequence-based PPI prediction methods utilize AA features such as hydrophobicity, physiochemical properties, evolutionary profiles, AA composition, AA mean, or weighted average over a sliding window [71]. Sequence-based methods can be categorized into statistical and Machine Learning (ML)-based methods. This section presents and evaluates some of the existing sequence-based approaches.

#### Statistical Sequence-Based Approaches

This section presents and describes several existing statistical sequence-based PPI prediction approaches.

##### **Mirror Tree Method:**

Pazos and Valencia [72] introduced the Mirror Tree Method based on the comparison of the evolutionary distances between the sequences of the associated protein families and using topological similarity of phylogenetic trees to predict PPI. These distances were calculated as the average value of the residue similarities taken from the McLachlan amino acid homology matrix [73]. The similarity between trees was calculated as the correlation between the distance matrices used to build the trees. The Mirror Tree Method does not require the creation of the phylogenetic trees but only the underlying distance matrices are analyzed, and therefore, this approach is independent of any given tree-construction method. Although the mirror tree method does not require the presence of fully sequenced genomes, it requires the presence of the orthologous proteins in all the species under consideration. As a result, when more species genomes become available, fewer proteins could be applied. In addition to that, the method is restricted to cases where at least eleven sequences were collected from the

same species for both proteins. This minimum limit was set empirically as a compromise between being sufficiently small to provide enough cases and large enough for the matrices to contain sufficient information. The approach can be improved by increasing the number of possible interactions by collecting sequences from a larger number of genomes. Further, since the distance matrices are not a perfect representation of the corresponding phylogenetic trees, it is possible that some inaccuracies are introduced by comparing distance matrices instead of the real phylogenetic trees.

#### **PIPE:**

Pitre *et al.* [74] introduced PIPE (Protein-protein Interaction Prediction Engine) to estimate the likelihood of interactions between pairs of the yeast *Saccharomyces cerevisiae* proteins using protein primary structure information. PIPE is based on the assumption that interactions between proteins occur by a finite number of short polypeptide sequences observed in a database of known interacting protein pairs. These sequences are typically shorter than the classical domains and reoccur in different proteins within the cell. PIPE estimates the likelihood of a PPI by measuring the reoccurrence of these short polypeptides within known interacting proteins pairs. To determine whether two proteins  $A$  and  $B$  interact, the two query proteins are scanned for similarity to a database of known interacting proteins pairs. For each known interacting pair  $(X, Y)$ , PIPE uses sliding windows to compare the AA residues in protein  $A$  against that in  $X$  and protein  $B$  against  $Y$ , and then measures how many times a window of protein  $A$  finds a match in  $X$  and at the same time a window in protein  $B$  matches a window in  $Y$ . These matches are counted and added up in a 2D matrix. A positive protein interaction is predicted when the reoccurrence count in certain cells of the matrix exceed a predefined threshold value. PIPE was evaluated on a randomly selected

set of 100 interacting yeast protein pairs and 100 non-interacting proteins from the database of interacting proteins (DIP) (<http://dip.doe-mbi.ucla.edu>) [75] and MIPS [76] databases. PIPE showed a prediction sensitivity of 0.61 and specificity of 0.89. Since PIPE is based on protein primary structure information without any previous knowledge about the higher structure, domain composition, evolutionary conservation or the function of the target proteins. It can identify interactions of protein pairs for which limited structural information is available. The limitations of PIPE are as follows. PIPE is computationally intensive and requires hours of computation per protein pair as it scans the interaction library repeatedly every time. Second, PIPE shows weakness in detecting novel interactions among genome wide large-scale datasets as it reported a large number of false positives. Third, PIPE was evaluated on uncertain data of interactions that were determined using several methods, each having a limited accuracy.

Pitre *et al.* [77] then developed PIPE2 as an improved and more efficient version of PIPE which showed a specificity of 0.999. PIPE2 represents AA sequences in a binary code which speeds up searching the similarity matrix. Unlike the original PIPE that scans the interaction database repeatedly every time, PIPE2 pre-computes all window comparisons in advance and stores them on a local disk.

Although PIPE2 achieves a high specificity, it has a large number of false positives with a sensitivity of 0.146 only. False positives rate can be reduced by incorporating other information about the target protein pairs including sub-cellular localization or functional annotation. A major limitation of PIPE2 is that it relies exclusively on a database of pre-existing interaction pairs for the identification of re-occurring short polypeptide sequences and in the absence of sufficient data, PIPE2 will be ineffective. PIPE2 is also less effective for motifs that span discontinuous primary sequence as it

does not account for gaps within the short polypeptide sequences.

### **Co-evolutionary Divergence:**

Liu *et al.* [78] introduced a sequence-based co-evolution PPI prediction method in the human proteins. The authors defined the co-evolutionary divergence (CD) based on two assumptions. First, PPI pairs may have similar substitution rates. Second, protein interaction is more likely to conserve across related species. CD is defined as the absolute value of the substitution rate difference between two proteins. CD can be used to predict PPI as the CD values of interacting protein pairs are expected to be smaller than those of non-interacting pairs. The method was evaluated using 172,338 protein sequences obtained from Evola database [79] for Homo sapiens and their orthologous protein sequences in thirteen different vertebrates. The PPI dataset was downloaded from the Human Protein Reference Database [80]. Pairwise alignment of the orthologous proteins was made with ClustalW2 software. The absolute value of substitution rate difference between two proteins was used to measure the CDs of protein pairs which were then used to construct the likelihood ratio table of interacting protein pairs.

The CD method combines co-evolutionary information of interacting protein pairs from many species. The method does not use multiple alignments, thus taking less time than other alignment methods such as the mirror tree method. The method is not limited to proteins with orthologous across all species under consideration. However, increasing the number of species will provide more information to improve the accuracy of the co-evolutionary divergence method. Although this method could rank the likelihood of interaction for a given pair of proteins, it did not infer specific features of interaction such as the interacting residues in the interfaces.

<b>Approach</b>	<b>Extracted Features</b>	<b>Technique/Tool</b>	<b>Datasets</b>
Mirror Tree (Pazos and Valencia 2001)	Similarity of phylogenetic trees	Evolutionary distance, McLachlan AA homology matrix	<i>Escherichia coli</i> protein (Dandekar <i>et al.</i> 1998)
PIPE (Pitre <i>et al.</i> 2006, 2008)	Short AA polypeptides	Similarity measure	Yeast protein (DIP and MIPS)
Co-evolutionary Divergence (Liu <i>et al.</i> 2013)	Co-evolutionary information,	Pairwise alignment, ClustalW2	Human protein (Matsuya <i>et al.</i> 2008, Prasad <i>et al.</i> 2009)

Table 2.3: Statistical Sequence-based PPI prediction approaches.

Table 2.3 summarizes these statistical sequence-based approaches including the features that are used, the technique and/or the tools applied, and the validation datasets used.

### **Machine-learning sequence-based PPI prediction approaches.**

This section describes several existing ML sequence-based PPI prediction approaches.

#### **Auto Covariance:**

Guo *et al.* [81] proposed a sequence-based method using Auto Covariance (AC) and Support Vector Machines (SVM). AA residues were represented by seven physicochemical properties. These properties are hydrophobicity, hydrophilicity, volumes of side chains, polarity, polarizability, solvent-accessible surface area, and net charge index of AA side chains. AC counts for the interactions between residues a certain distance apart in the sequence. AA physicochemical properties were analyzed by AC based on the calculation of covariance. A protein sequence was characterized by a series of ACs that covered the information of interactions between each AA residue and its 30 vicinal residues in the sequence. Finally, a SVM model with a Radial Basis Function (RBF) kernel was constructed using the vectors of AC variables as input. The optimization

experiment demonstrated that the interactions of one AA residue and its 30 vicinal AAs would contribute to characterizing the PPI information. The software and datasets are available at [http://www.scucic.cn/Predict\\_PPI/index.htm](http://www.scucic.cn/Predict_PPI/index.htm). A dataset of 11,474 yeast PPI pairs extracted from DIP [82] was used to evaluate the model and the average prediction accuracy, sensitivity, and precision achieved are respectively 0.86, 0.85, and 0.87.

One of the advantages of this approach is that AC includes long-range interaction information of AA residues which are important in PPI identification. The use of SVM as a predictor is another advantage. SVM is the state of the art ML technique and has many benefits and overcomes many limitations of other techniques. SVM has strong foundations in statistical learning theory [83] and has been successfully applied in various classification problems [84]. SVM offers several related computational advantages such as the lack of local minima in the optimization [85].

### **Pairwise Similarity:**

Zaki *et al.* [14] proposed a PPI predictor based on pairwise similarity of protein primary structure. Each protein sequence was represented by a vector of pairwise similarities against large AA subsequences created by a sliding window which passes over concatenated protein training sequences. Each coordinate of this vector is the  $E$ -value of the Smith-Waterman (SW) score [86]. These vectors were then used to compute the kernel matrix which was exploited in conjunction with a RBF-kernel SVM. Two proteins may interact by the means of the scores similarities they produce [87, 88]. Each sequence in the testing set was aligned against each sequence in the training set, counted the number of positions that have identical residues, and then divided by the total length of the alignment.

The method was evaluated on a dataset of yeast *Saccharomyces cerevisiae* proteins created by Chen and Liu [89] and contains 4917 interacting protein pairs and 4000 non-interacting pairs. The method achieved an accuracy of 0.78, a sensitivity of 0.81, a specificity of 0.744, and a ROC of 0.85.

SW alignment score provides a relevant measure of similarity between proteins. Therefore protein sequence similarity typically implies homology, which in turn may imply structural and functional similarity [90]. SW scores parameters have been optimized over the past two decades to provide relevant measures of similarity between sequences and they now represent core tools in computational biology [91]. The use of SVM as a predictor is another advantage. This work can be improved by combining knowledge about gene ontology, inter-domain linker regions, and interacting sites to achieve more accurate prediction.

### **AA Composition:**

Roy *et al.* [92] examined the role of amino acid composition (AAC) in PPI prediction and its performance against well-known features such as domains, tuple feature, and signature product feature. Every protein pair was represented by AAC and domain features. AAC was represented by monomer and dimer features. Monomer features capture composition of individual amino acids, whereas dimer features capture composition of pairs of consecutive AAs. To generate the monomer features, a 20-dimensional vector representing the normalized proportion of the 20 AAs in a protein was created. The real-valued composition was then discretized into 25 bits producing a set of 500 binary features. To generate the dimer features, a 400-dimensional vector of all possible AA pairs were extracted from the protein sequence and discretized into 10 bits producing a set of 4000 binary features. The domains were represented as binary

features with each feature identified by a domain name. To compare AAC against other non-domain sequence-based features, tuple features [93] and signature products [94] were obtained. The tuple features were created by grouping AAs into six categories based on their biochemical properties, and then creating all possible strings of length 4 using these categories. The signature products were obtained by first extracting signatures of length 3 from the individual protein sequences. Each signature consists of a middle letter and two flanking AAs represented in alphabetical order. Thus two 3-tuples with the first and third amino acid letter permuted have the same signature. The signatures were used to construct a signature kernel specifying the inner product between two proteins.

The proposed approach was examined using three machine learning classifiers (logistic regression, SVM, and the Naive Bayes) on PPI datasets from yeast, worm and fly. Three datasets for yeast *S. cerevisiae* were extracted from the General Repository for Interaction Datasets (GRID) database [95], TWOHYB (Yeast Two-hybrid), AFFMS (Affinity pull down with mass spectrometry), and PCA (protein complementation assay). In addition to that, a dataset each for worm, *C. elegans* (Biogrid dataset) [96] and fly, *D. melanogaster* [95] were used. The authors reported that AAC features performed almost equivalent contribution as domain knowledge across different datasets and classifiers which indicated that AAC captures significant information for identifying PPI. AAC is a simple feature, computationally cheap, applicable to any protein sequence, and can be used when there is lack of domain information. AAC can be combined with other features to enhance PPI prediction.

#### **AA Triad:**

Yu *et al.* [97] proposed a probability-based approach of estimating triad signif-

icance to alleviate the effect of AA distribution in nature. The relaxed variable kernel density estimator (RVKDE) [98] was employed to predict PPI based on AA triad information. The method is summarized as follows. Each protein sequence was represented as AA triads by considering every three continuous residues in the protein sequence as a unit. To reduce feature dimensionality vector, the 20 AA types were categorized into seven groups based on their dipole strength and side chain volumes [69]. The triads were then scanned one by one along the sequence, and each scanned triad is counted in an occurrence vector,  $O$ . Subsequently, a significance vector,  $S$ , was proposed to represent a protein sequence by estimating the probability of observing less occurrences of each triad than the one that is actually observed in  $O$ . Each PPI pair was then encoded as a feature vector by concatenating the two significance vectors of the two individual proteins. Finally, the feature vector was used to train a RVKDE PPI predictor. The method was evaluated on 37,044 interacting pairs within 9,441 proteins from the Human Protein Reference Database (HPRD) [99, 100]. Datasets with different positive-to-negative ratios (from 1:1 to 1:15) were generated with the same positive instances and distinct negative sets, which are obtained by randomly sampling from the negative instances. The authors concluded that the degree of dataset imbalance is important to PPI predictor behavior. With 1:1 positive-to-negative ratio, the proposed method achieves 0.81 sensitivity, 0.79 specificity, 0.79 precision, and 0.8 F-measure. These evaluation measures drop as the data gets more imbalanced to reach 0.39 sensitivity, 0.97 specificity, 0.495 precision, and 0.44 F-measure with 1:15 positive-to-negative ratio.

RVKDE is a ML algorithm that constructs a RBF neural network to approximate the probability density function of each class of objects in the training dataset. One main distinct feature of RVKDE is that it takes an average time complexity of  $O(n \log n)$  for the model training process, where  $n$  is the number of instances in the

training set. In order to improve the prediction efficiency, RVKDE considers only a limited number of nearest instances within the training dataset to compute the kernel density estimator of each class. One important advantage of RVKDE, in comparison with SVM, is that the learning algorithm generally takes far less training time with an optimized parameter setting. In addition to that, the number of training samples remaining after a data reduction mechanism is applied is quite close to the number of support vectors of SVM algorithm. Unlike SVM, RVKDE is capable of classifying data with more than two classes in one single run [98].

#### **UNISPPPI:**

Valente *et al.* [101] (2013) introduced UNISPPPI (Universal In Silico Predictor of Protein-Protein Interactions). The authors examined both the frequency and composition of the physicochemical properties of the twenty protein AAs to train a decision tree PPI classifier. The frequency feature set includes the percentages of each of the 20 AA in the protein sequence. The composition feature set was obtained by grouping each AA of a protein into one of three different groups related to seven physicochemical properties and calculating the percentage of each group for each feature ending up by a total of 21 composition features. The seven physicochemical properties are hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure, and solvent accessibility. When tested on a dataset of PPI pairs of twenty different eukaryotic species including eukaryotes, prokaryotes, viruses, and parasite-host associations, UNISPPPI correctly classified 0.79 of known PPI pairs and 0.73 of non-PPI pairs. The authors concluded that using only the AA frequencies was sufficient to predict PPIs. They further concluded that the AA frequencies of Asparagines (N), Cysteine (C), and Isoleucine (I) are important features for distinguishing between interacting

and non-interacting protein pairs.

The main advantages of UNISPPI are its simplicity and low computational cost as small amount of features were used to train the decision tree classifier. Decision tree classifier is fast to build and has few parameters to tune. Decision trees can be easily analyzed and the features can be ranked according to their capabilities of distinguishing PPIs from non-PPIs. However, decision tree classifiers normally suffer from overfitting.

**ETB-Viterbi:**

Kern *et al.* [102] proposed the Early Traceback Viterbi (ETB-Viterbi) as a decoding algorithm with an early traceback mechanism in ipHMMs (Interaction Profile Hidden Markov Models) [103] which was designed to optimally incorporate long-distance correlations between interacting AA residues in input sequences. The method was evaluated on real data from the 3DID database [104] along with simulated data generated from 3DID data containing different degrees of correlation and reversed sequence orientation. ETB-Viterbi was capable to capture the long-distance correlations for improved prediction accuracy and was not much affected by sequence orientation. Hidden Markov models (HMM) are powerful probabilistic modeling tool for analyzing and simulating sequences of symbols that are emitted from underlying states and not directly observable [105]. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states. However, the Viterbi algorithm is expensive in terms of memory and computing time. The HMM training involves repeated iterations of the Viterbi algorithm which makes it quite slow. HMM Model may not converge to a truly optimal parameter set for a given training set as it can be trapped in local maxima, and can suffer from overfitting [106, 107, 108, 109].

Table 2.4 summarizes these ML sequence-based approaches and compared them in terms of features, techniques, tools, and validation datasets.

**2.3.2 Structure-Based Approaches**

Structure-based PPI prediction methods use three-dimensional structural features such as domain information, solvent accessibility, secondary structure states, and hydrophobic and polar surface locations [71]. Structure-based PPI prediction methods can be categorized into template-based, statistical, and ML-based methods. This sec-

Approach	Extracted Features	Technique/Tool	Datasets
Auto Covariance (Guo <i>et al.</i> 2008)	AA physicochemical properties	Auto covariance, SVM	Yeast protein (DIP and MIPS)
Pairwise Similarity (Zaki <i>et al.</i> 2009)	Pairwise similarity	SVM	Yeast protein
AA Composition (Roy <i>et al.</i> 2009)	AAC	Logistic regression, SVM, Naive Bayes	Yeast protein, worm protein, fly protein
AA Triad (Yu <i>et al.</i> 2010)	AA triad information	RVKDE	Human protein (HPRD)
UNISPPI (Valente <i>et al.</i> 2013)	Frequency and composition of AA physiochemical properties	Decision trees	Twenty different eukaryotic species
ETB-Viterbi (Kern <i>et al.</i> 2013)	AA residues	HMM, Early Traceback Viterbi	3DID database

Table 2.4: Machine-learning sequence-based PPI prediction approaches.

tion presents and evaluates some of the state-of-the-art structure-based approaches.

### Template Structure-Based Approaches

Examples of template structure-based approaches are PRISM and PrePPI.

#### PRISM:

Tuncbag *et al.* [110] developed PRISM as a template-based PPI prediction method based on information regarding the interaction surface of crystalline complex structures. The two sides of a template interface are compared with the surfaces of two target monomers by structural alignment. If regions of the target surfaces are similar to the complementary sides of the template interface, then these two targets are predicted to interact with each other through the template interface architecture. The method can be summarized as follows. First, interacting surface residues of target chains are extracted using Naccess [111]. Second, complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt

[112]. Third, the structural alignment results are filtered according to threshold values, and the resulting set of target surfaces is transformed into the corresponding template interfaces to form a complex. Finally, the Fiber-Dock [113] algorithm is used to refine the interactions to introduce flexibility, compute the global energy of the complex, and rank the solutions according to their energies. When the computed energy of a protein pair is less than a threshold of -10 kcal/mol, the pair is determined to interact.

PRISM has been applied for predicting PPI in a human apoptosis pathway [114] and a p53- protein-related pathway [115], and has contributed to the understanding of the structural mechanisms underlying some types of signal transduction. PRISM obtained a precision of 0.231 when applied to a human apoptosis pathway that consisted of 57 proteins.

### **PrePPI:**

Zhang *et al.* [28] proposed PrePPI (Predicting Protein-Protein Interactions) as a structural alignment PPI predictor based on geometric relationships between secondary structure information. Given a pair of query proteins  $A$  and  $B$ , representative structures for the individual subunits ( $M_A$ ,  $M_B$ ) are taken from the PDB (Protein Data Bank) [116] or from the ModBase [117] and SkyBase [118] homology model databases. Close and remote structural neighbors are found for each subunit. A template for the interaction exists if a PDB or PQS [119] structure contains interacting pairs that are structural neighbors of  $M_A$  and  $M_B$ . A model is constructed by superposing the individual subunits,  $M_A$  and  $M_B$ , on their corresponding structural neighbors. The likelihood for each model to represent a true interaction is then calculated using a Bayesian Network trained on 11,851 yeast interactions and 7,409 human interactions datasets. Finally the structure-derived score is combined with non-structural informa-

tion, including co-expression and functional similarity, into a naive Bayes classifier.

Although template-based methods can achieve high prediction accuracy when close templates are retrieved, the accuracy significantly decreases when the sequence identity of target and template is low.

## **Statistical Structure-Based Approaches**

This section describes several existing statistical structure-based PPI prediction approaches.

### **PID Matrix Score:**

Kim *et al.* [13] presented the Potentially Interacting Domain pair (PID) matrix as a domain-based PPI prediction algorithm. The PID matrix score was constructed as a measure of interactability (interaction probability) between domains. The algorithm analysis was based on the DIP (Database of Interacting Proteins) which contains more than ten thousand of mostly experimentally verified interacting protein pairs. Domain information was extracted from InterPro [120] which is an integrated database of protein families, domains and functional sites. Cross validation was performed with subsets of DIP data (positive datasets) and randomly generated protein pairs from TrEMBL/SwissProt database (negative datasets). The method achieved 0.50 sensitivity and 0.98 specificity. The authors reported that the PID matrix can also be used in the mapping of the genome-wide interaction networks.

### **PreSPI:**

Han *et al.* [121, 122] proposed a domain combination-based method which considers all possible domain combinations as the basic units of protein interactions. The

domain combination interaction probability is based on the number of interacting protein pairs containing the domain combination pair and the number of domain combinations in each protein. The method considers the possibility of domain combinations appearing in both interacting and non-interacting sets of protein pairs. The ranking of multiple protein pairs were decided by the interacting probabilities computed through the interacting probability equation.

The method was evaluated using an interacting set of protein pairs in yeast acquired from DIP database [75], and a randomly generated non-interacting set of protein pairs. The domain information for the proteins was extracted from the PDB<sup>6</sup> [120, 116]. PreSPI achieved a sensitivity of 0.77 and a specificity of 0.95.

PreSPI suffers from several limitations. First, this method ignores other domain-domain interaction information between the protein pairs. Second, it assumes that one domain combination is independent of another. Third, the method is computationally expensive as all possible domain combinations are considered.

### **Domain Cohesion and Coupling:**

Jang *et al.* [123] proposed a domain cohesion and coupling (DCC)-based PPI prediction method using the information of intra-protein domain interactions and inter-protein domain interactions. The method aims to identify which domains are involved in a PPI by determining the probability of the domains causing the proteins to interact irrespective of the number of participating domains. The coupling powers of all domain interaction pairs are stored in an interaction significance (IS) matrix which is used to predict PPI. The method was evaluated on *S. cerevisiae* proteins and achieved 0.82 sensitivity and 0.83 specificity. The domain information for the proteins was extracted

---

<sup>6</sup><http://www.ebi.ac.uk/proteome/>

from Pfam (<http://pfam.sanger.ac.uk>) [49], which is a protein domain family database that contains multiple sequence alignments of common domain families.

### **MEGADOCK:**

Ohue *et al.* [124] developed MEGADOCK as a protein-protein docking software package using the real Pairwise Shape Complementarity (rPSC) score. First, they conducted rigid-body docking calculations based on a simplified energy function considering shape complementarities, electrostatics, and hydrophobic interactions for all possible binary combinations of proteins in the target set. Using this process, a group of high-scoring docking complexes for each pair of proteins were obtained. Then, ZRANK [125] was applied for more advanced binding energy calculation and re-ranked the docking results based on ZRANK energy scores. The deviation of the selected docking scores from the score distribution of high-ranked complexes was determined as a standardized score (Z-score) and was used to assess possible interactions. Potential complexes that had no other high-scoring interactions nearby were rejected using structural differences. Thus binding pairs that had at least one populated area of high-scoring structures were considered. MEGADOCK has been applied for PPI prediction for 13 proteins of a bacterial chemotaxis pathway [126, 127] and obtained a precision of 0.4. MEGADOCK is available at <http://www.bi.cs.titech.ac.jp/megadock>.

One of the limitations of this approach is the demerit of generating false-positives for the cases in which no similar structures are seen in known complex structure databases.

### **Meta Approach:**

Ohue *et al.* [128] proposed a PPI prediction approach based on combining

template-based and docking methods. The approach applies PRISM [110] as a template-matching method and MEGADOCK [124] as a docking method. A protein pair is considered to be interacting if both PRISM and MEGADOCK predict that this protein pair interacts. When applied to the human apoptosis signaling pathway, the method obtained a precision of 0.333, which is higher than that achieved using individual methods (0.231 for PRISM and 0.145 for MEGADOCK), while maintaining an F1 of 0.285 comparable to that obtained using individual methods (0.296 for PRISM, and 0.220 for MEGADOCK).

Meta approaches have already been used in the field of protein tertiary structure prediction [129], and critical experiments have demonstrated improved performance of Meta predictors when compared with individual methods. The Meta approach has also provided favorable results in protein domain prediction [53] and the prediction of disordered regions in proteins [130]. Although some true positives may be dropped by this method, the remaining predicted pairs are expected to have higher reliability because of the consensus between two prediction methods that have different characteristics.

## Machine Learning Structure-Based Approaches

Examples of ML structure-based approaches are Maximum Likelihood Estimation [131], Random Forest [89], and Struct2Net.

### MLE:

Deng *et al.* [131] developed the Maximum Likelihood Estimation (MLE) method which is based on the assumption that two proteins interact if at least one pair of domains of the two proteins interact. It infers domain interactions by maximizing the likelihood of the observed protein interaction data. The probabilities of interaction

between two domains (only single-domain pair is considered) are optimized using the expectation-maximization (EM) algorithm. They used a combined interaction data which was experimentally obtained through two hybrid assays on *Saccharomyces cerevisiae* by Uetz *et al.* [132] and Ito *et al.* [133]. The protein domain information were collected from Pfam database [134].

The basic assumptions of this method ignore the following biological factors. First, the method assumes independence of domain-domain interactions. However, the fact that two domains interact or not may depend on other domains in the same protein or other environmental conditions. Second, although the method identified domains that coexist in proteins and merged them as one domain, there certainly exist many domains whose functions depend on other domains in the same protein. Third, the idea of using domain-domain interactions to predict protein-protein interactions assumes that some subunits with special structure are essential to protein-protein interactions. These subunits may be different from PFAM domains obtained through multiple alignments. Fourth, the method used PFAM-B domains in the same level as the PFAM-A domains. However, PFAM-B domains are shorter and less known than PFAM-A domains, and therefore, their roles in protein-protein interactions may not be the same.

### **Random Forest:**

Chen and Liu [89] introduced a domain-based Random Forest PPI predictor. Protein pairs were characterized by the domains existing in each protein. The protein domain information were collected from Pfam database [134]. Each protein pair was represented by a vector of features where each feature corresponds to a Pfam domain. If a domain exists in both proteins, then the associated feature value is 2. If the domain exists in one of the two proteins, then its associated feature value is 1. If a domain does

not exists in both proteins, then the feature value is 0. These domain features were used to train a Random Forest PPI classifier. The random decision forest constructs many decision trees and each is grown from a different subset of training samples and random subset of feature and the final classification of a given protein pair is determined by majority votes among the classes decided by the forest of trees.

When evaluated on a dataset containing 9834 yeast protein interaction pairs among 3713 proteins, and 8000 negative randomly generated samples, the method achieved a sensitivity of 0.8 and a specificity of 0.64.. Yeast PPI data was collected from the DIP [75, 82], Deng *et al.* [131], Schwikowski *et al.* [135]. The dataset of Deng *et al.* is a combined interaction data experimentally obtained through two hybrid assays on *Saccharomyces cerevisiae* by Uetz *et al.* [132] and Ito *et al.* [133]. Schwikowski *et al.* gathered their data from yeast two-hybrid, biochemical and genetic data.

Random Forest classifier has several advantages. It is relatively fast, simple, robust to outliers and noise, easily parallelized, avoids overfitting, and performs well in many classification problems [136, 137]. Random Forest shows a significant performance improvement over the single tree classifiers. It interprets the importance of the features using measures such as decrease mean accuracy or *Gini* importance [138]. RF benefit from the randomization of decision trees as they have low-bias and high variance. Random Forest has few parameters to tune and less dependent on tuning parameters [139, 140]. However, the computational cost of Random Forest increases as the number of generated trees increases. One of the limitations of this approach is that PPI prediction depends on domain knowledge so proteins without domain information cannot provide any useful information for prediction. Therefore, the method excluded the pairs where at least one of the proteins has no domain information.

**Struct2Net:**

Singh *et al.* [141] introduced Struct2Net as a structure-based PPI predictor. The method predicts interactions by threading each pair of protein sequences into potential structures in the Protein Data Bank (PDB) [116]. Given two protein sequences (or one sequence against all sequences of a species), Struct2Net threads the sequence to all the protein complexes in the PDB and then chooses the best potential match. Based on this match, it uses logistic regression technique to predict whether the two proteins interact.

Later on, Singh *et al.* [142] introduced Struct2Net as a web server with multiple querying options which is available at <http://struct2net.csail.mit.edu>. Users can retrieve Yeast, fly, and human PPI predictions by gene name or identifier while they can query for proteins of other organisms by AA sequence in FASTA format. Struct2Net returns a list of interacting proteins if one protein sequence is provided and an interaction prediction if two sequences are provided. When evaluated on yeast and fly protein pairs, Struct2Net achieves a recall of 0.80 with a precision of 0.30.

A common limitation of all structure-based PPI prediction approaches is the low coverage as the number of known protein structures is much smaller than the number of known protein sequences, and therefore, such approaches fail when there is no structural template available for the queried protein pair. Table 2.5 summarizes these structure-based approaches and compared them in terms of features, techniques, tools, and validation datasets.

Several approaches for predicting interactions between human and HIV proteins have been proposed. Tastan *et al.* [143] proposed a random forest classification model for predicting HIV-1-human PPI. Dyer *et al.* [144] proposed a SVM-based approach for predicting physical interactions between human and HIV proteins. Mukhopadhyay

Approach	Extracted Features	Technique/ Tool	Datasets
PRISM (Tuncbag <i>et al.</i> 2011)	Interaction surface of crystalline complex structures	Naccess, MultiProt, Fiber-Dock	Human Protein (Ozbabacan <i>et al.</i> 2012, Tuncbag <i>et al.</i> 2009)
PrePPI (Zhang <i>et al.</i> 2012)	Secondary structure	Bayesian networks, Naive Bayes	Yeast protein, Human protein
PID Matrix Score (Kim <i>et al.</i> 2002)	Potentially interacting domain pairs	PID matrix	DIP, InterPro, TrEMBL/SwissProt
PreSPI (Han <i>et al.</i> 2003, 2004)	Domain combination interaction probability	Interacting probability equation	Yeast protein (DIP), PDB
DCC (Jang <i>et al.</i> 2012)	Intra-protein and inter-protein domain interactions	Interaction significance matrix	<i>S. cerevisiae</i> protein, Pfam
MEGADOCK (Ohue <i>et al.</i> 2013a)	Shape complementaries, electrostatics, and hydrophobic interactions	rPSC, ZRANK	Bacterial protein (Ohue <i>et al.</i> 2012, Matsuzaki <i>et al.</i> 2013)
Meta Approach (Ohue <i>et al.</i> 2013b)	Interaction surface of crystalline complex structures, shape complementaries, electrostatics, and hydrophobic interactions	PRISM, MEGADOCK	Human protein
MLE (Deng <i>et al.</i> 2002)	Interacting domains	Maximum Likelihood Estimation	Uetz <i>et al.</i> , Ito <i>et al.</i> , Pfam
Random Forest (Chen and Liu 2005)	Existence of similar domains	Random Forest	DIP, Deng <i>et al.</i> , Schwikowski <i>et al.</i> , Pfam
Struct2Net (Singh <i>et al.</i> 2006, 2010)	Homology with known protein complexes in PDB	Logistic regression	Yeast, Fly ,and Human protein

Table 2.5: Structure-based PPI prediction approaches.

*et al.* [145] proposed an association rule mining technique for discovering a set of rules among human and HIV-1 proteins.

Most of the discussed PPI prediction methods have the following limitations:

- They are based on previously identified domains, and therefore they cannot be applied when domain knowledge is not available.
- Although protein domains are highly informative for PPI prediction, other se-

quence parts such as linkers can also significantly contribute to PPI prediction.

- They have, in general, limited capabilities to detect novel interactions and to differentiate them from false positives [146, 14].

In this work, I develop a compact and accurate approach that integrates domain-linker prediction with PPI prediction based solely on protein primary structure information. This is achieved through introducing the concept of amino acid (AA) compositional index. The compositional index is deduced from the protein sequence dataset of domain-linker segments. The compositional index is then combined with physiochemical properties to construct a novel AA profile. A sliding window of variable length is used to extract the information on the dependencies of each AA and its neighboring residues. The extracted information is then used to train a machine-learning classifier to predict novel domains and linkers. Once domains are identified within proteins, protein interactions can be predicted by analyzing their interacting domains. The proposed approach efficiently processes high-dimensional multi-domain protein data with a more accurate predictive performance than existing state-of-the-art approaches.

## Chapter 3: Research Methodology

This chapter provides an overview of the research method in Section 3.1, describes the datasets in Section 3.2 and defines the evaluation measures in Section 3.3

### 3.1 Method Overview

In this work, I develop a compact and accurate approach that integrates structural domain and inter-domain linker prediction with PPI prediction based solely on protein primary structure information. The approach consists of two main stages; identifying structural domains within protein sequences and predicting PPI. The first stage includes two main contributions. The first contribution is predicting inter-domain linker regions by introducing the concept of AA compositional index and refining the prediction using Simulated Annealing. The compositional index of an amino acid represents the preference of this AA to appear in linker regions based on its frequencies in linker and domain regions. The second contribution is identifying structural domains based on inter-domain linker knowledge by constructing a protein profile that combines amino acid compositional index and physiochemical properties and developing a machine-learning classifier for predicting novel domains and linkers. In the second stage we predict PPIs by characterizing structural domains within proteins and analyzing their domain-domain interactions. An overview of the method is illustrated in Figure 3.1.

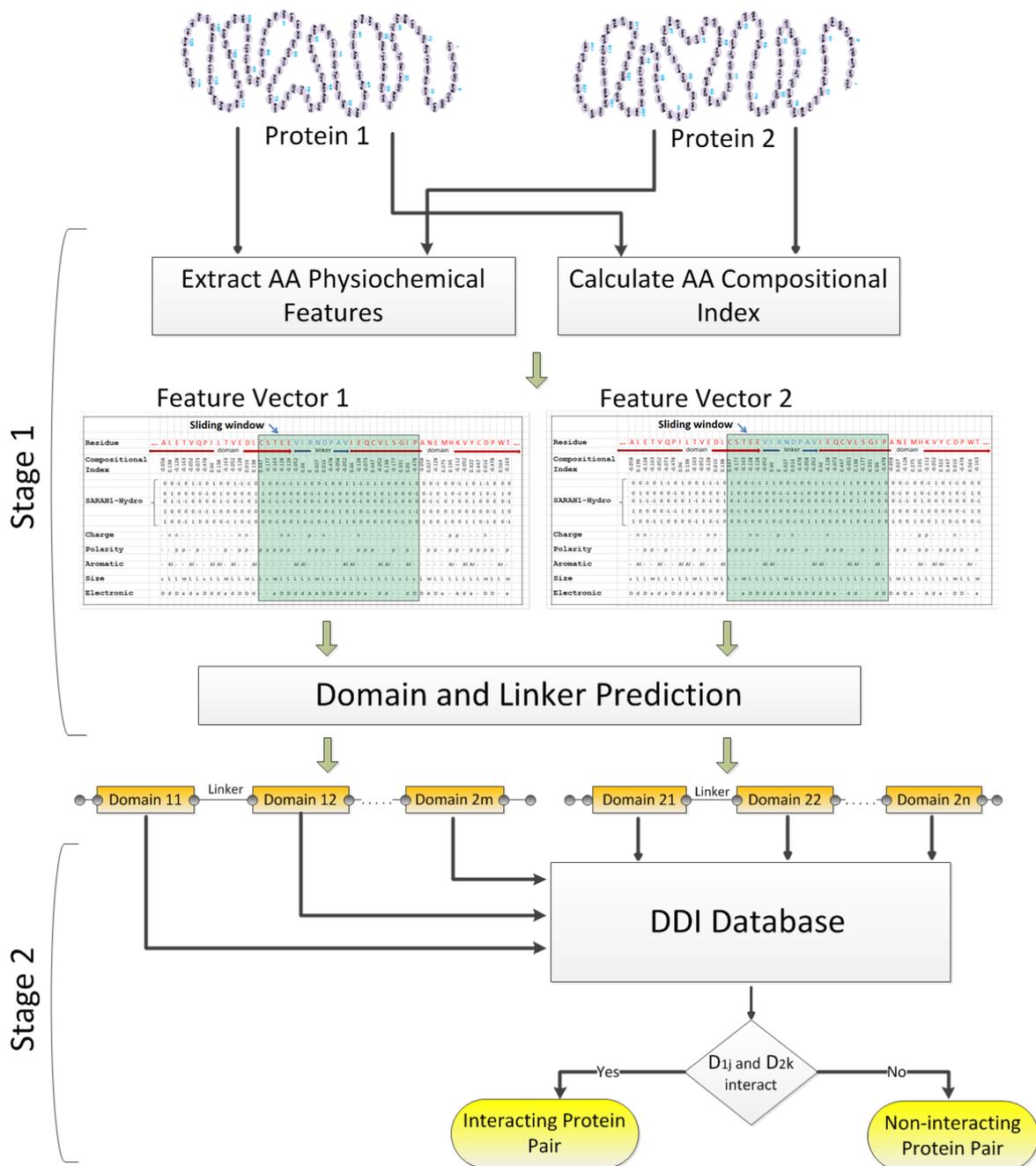


Figure 3.1: Method overview.

The two main stages of this work, which are aligned to our main objectives, can be summarized as follows:

- Developing a novel method for identifying structural domains within protein sequences. This is achieved through the following steps:
  - (1) Predicting protein domain-linker regions by introducing the concept of AA compositional index and refining the prediction by Simulated Annealing.
  - (2) Developing a machine-learning approach for predicting novel domains and linkers:
    - (i) To include more biological knowledge, the compositional index is combined with AA physiochemical properties to construct a protein profile.
    - (ii) A sliding window technique is applied to extract the information on the dependencies of each AA and its neighbors.
    - (iii) A Random Forest classifier is developed to distinguish between domains and inter-domain linker regions.
- Developing a novel PPI predictor:
  - (1) Characterizing structural domains within protein sequences.
  - (2) Identifying interacting domains.
  - (3) Predicting protein interactions based on analyzing their interacting domains.

To evaluate the performance of our proposed method and to compare our experimental results with other approaches, we used benchmark datasets along with standard evaluation measures. These datasets and evaluation measures are described in following sections.

## 3.2 Datasets

### 3.2.1 Structural Domains and Inter-Domain Linker Prediction

To evaluate the performance of the inter-domain linker prediction and structural domain prediction approach, two protein sequence datasets were used. The first dataset is DS-All [6, 7] which was used to evaluate DROP [31]. All the sequences in DS-All were extracted from the non-redundant Protein Data Bank (nr-PDB) chain set<sup>1</sup> and contains 182 protein sequences including 216 linker segments. By examining each sequence carefully, we found that the assignment of domains in DS-All dataset is inconsistent with the ones in PDB. We thus validated the domain and inter-domain linkers according to NCBI conserved domains database<sup>2</sup> and ended up with 140 sequences including 334 domains and 183 linker segments. The average numbers of AA residues in linker segments is 12.7 with a standard deviation of 13.8 and the average numbers of AA residues in domain segments are 147.1 with a standard deviation of 90.1.

The protein sequences in the second set were extracted from the Swiss-Prot database [54] and have tested by Suyama and Ohara [4] to evaluate the performance of DomCut. This dataset contains 273 non-redundant protein sequences including 486 linker and 794 domain segments. The average numbers of AA residues in linker segments is 35.8 with a standard deviation of 26.7 and the average numbers of AA residues in domain segments are 122.1 with a standard deviation of 136.3. Therefore, about 85% (794 x 122.1) of the total AA residues exist in domain segments and only 15% (486 x 35.5) are in linker segments. The two datasets are summarized in Table 3.1.

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>

<sup>2</sup><http://www.ncbi.nlm.nih.gov/protein>

<b>Dataset</b>	<b>DS-All</b>	<b>DomCut/Swiss-Prot</b>
Reference	Ebina et al. [6, 7]	Suyama and Ohara [4]
Number of proteins	140	273
Number of linkers	183	486
Number of domains	334	794
Average number of AAs in linkers	13	36
Average number of AAs in domains	147	122

Table 3.1: Summary of domain-linker datasets.

Table 3.2 summarizes the protein resources and tools that we used in validating domain and linker prediction.

<b>Tool</b>	<b>Resource</b>	<b>Website</b>
PFam	The Protein family database	<a href="http://pfam.xfam.org/">http://pfam.xfam.org/</a>
NCBI	The National Center for Biotechnology Information	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
RCSB/PDB	Protein Data Bank	<a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a>

Table 3.2: Protein Tools.

### 3.2.2 PPI Prediction

To evaluate the performance of our PPI prediction approach, we used a dataset containing 4,917 yeast *Saccharomyces cerevisiae* protein interaction pairs among 3,713 proteins, and 4,000 negative randomly-generated samples. Yeast PPI data was collected from the DIP [75, 82], Deng *et al.* [131], Schwikowski *et al.* [135]. The dataset of Deng *et al.* is a combined interaction data experimentally obtained through two hybrid assays on *Saccharomyces cerevisiae* by Uetz *et al.* [132] and Ito *et al.* [133]. Schwikowski *et al.* gathered their data from yeast two-hybrid, biochemical and genetic data. As non-PPI data are unavailable, the negative samples were randomly generated. A protein pair is considered to be non-PPI if it does not exist in the interaction set. This dataset was

gathered and used by Chen and Liu [89]. Both the positive and negative PPI examples were divided evenly into training and testing datasets.

We obtained the domain information of the protein pairs from the Pfam-A release 27.0<sup>3</sup> [147] using the NCBI BLAST SOAP<sup>4</sup> [148, 51, 149] sequence similarity search tool.

To validate our PPI prediction, we used three Domain-Domain Interaction (DDI) databases; DOMINE, IDDI, and 3did. DOMINE<sup>5</sup> [150, 18] is a database of domain interactions inferred from experimentally characterized high-resolution 3D structures in the Protein Data Bank (PDB)<sup>6</sup>, in addition to predicted domain interactions by thirteen different computational approaches using Pfam domain definitions. DOMINE contains a total of 26,219 DDI pairs among 5,410 domains, out of which 6,634 are inferred from PDB entries, and 21,620 are predicted by at least one computational approach.

The integrated domain-domain interaction analysis system (IDDI)<sup>7</sup> [151] provides 20,4715 unique DDI pairs with different reliability scores. The reliability of the predicted DDI pairs are determined by considering the confidence score of the prediction method, the independence score of the predicted datasets, and the DDI prediction score measured by different prediction methods.

The database of 3D interacting domains (3did)<sup>8</sup> [152] is a collection of 3D structures of domain-based interactions in the PDB based on domain definitions from Pfam release 27.0 [147]. The 3did database contains 8,651 DDI pairs. Table 3.3 summarizes

---

<sup>3</sup><http://pfam.sanger.ac.uk>

<sup>4</sup>[http://www.ebi.ac.uk/Tools/webservices/services/sss/ncbi\\_blast\\_soap](http://www.ebi.ac.uk/Tools/webservices/services/sss/ncbi_blast_soap)

<sup>5</sup><http://domine.utdallas.edu>

<sup>6</sup><http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>

<sup>7</sup><http://pcode.kaist.ac.kr/iddi/>

<sup>8</sup><http://3did.irbbarcelona.org>

these DDI databases.

<b>Dataset</b>	<b>Number of DDI pairs</b>	<b>Website</b>
DOMINE	26,219	<a href="http://domine.utdallas.edu">http://domine.utdallas.edu</a>
IDDI	20,4715	<a href="http://pcode.kaist.ac.kr/iddi/">http://pcode.kaist.ac.kr/iddi/</a>
3did	8,651	<a href="http://3did.irbbarcelona.org">http://3did.irbbarcelona.org</a>

Table 3.3: DDI databases.

### 3.3 Evaluation Measures

The most commonly used evaluation metrics in general classification tasks are accuracy ( $Ac$ ), recall ( $R$ ), precision ( $P$ ), specificity ( $Sp$ ), F-measure, and Receiver Operating Characteristic (ROC).

$$Ac = \frac{TP + TN}{TP + TN + FN + FP} \quad (3.1)$$

$$R = \frac{TP}{TP + FN} \quad (3.2)$$

$$P = \frac{TP}{TP + FP} \quad (3.3)$$

$$Sp = \frac{TN}{TN + FP} \quad (3.4)$$

where  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  represent true positive, true negative, false positive, and false negative, respectively.

The F-measure (F1) is an evaluation metric that combines precision and recall into a single value. It is defined as the harmonic mean of precision and recall [153, 154]:

$$F1 = \frac{2PR}{P + R} \quad (3.5)$$

The Receiver Operating Characteristic (ROC) is a graphical plot that illustrates the classifier performance. The curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The ROC curve is thus the sensitivity as a function of false positive rate. Each prediction

result or instance represents one point in the ROC space. The best possible prediction method would yield a point in the upper left corner of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). Classifier accuracy is measured by the area under the ROC curve (AUC), and therefore, AUC is used in model comparison. An area of 1 represents a perfect test while an area of 0.5 represents a worthless test. [155].

We used recall, precision, F-measure, and AUC to evaluate our first and second contributions of domain and linker prediction approaches. Our third contribution is evaluated and compared with existing PPI prediction approaches using sensitivity (recall) and specificity.

In the proceeding chapters the proposed method will be discussed in details. Chapter 4 presents our first contribution in domain-linker prediction using AA compositional index and Simulated Annealing. Section 4.1 introduces the proposed formula for AA compositional index. Section 4.2 describes the use of Simulated Annealing algorithm to refine the domain-linker prediction by detecting the optimal threshold values of AA compositional index. Chapter 5 presents our second contribution in developing a machine-learning approach for predicting novel domains and linkers. Chapter 6 presents our third contribution which is predicting protein-protein interactions based on their identified domains.

## Chapter 4: CISA: Inter-Domain Linker Prediction Using Compositional Index and Simulated Annealing

In this chapter, we introduce our approach for predicting domain-linker regions using AA Compositional Index and Simulated Annealing which we call it CISA. CISA consists of two main steps; calculating the AA compositional index (CI) for the protein sequence of interest and then applying the simulated Annealing (SA) algorithm to refine the prediction by detecting the optimal set of threshold values that distinguish between domains and linker regions. In the first step, linker and domain segments are extracted from the protein sequence dataset and the frequencies of AA appearances in linker segments and non-linker segments are computed. Then, the AA composition of the query protein sequence is computed, and finally the AA compositional index is calculated. In the second step, SA is applied to find the optimal set of threshold values that separate linker segments from non-linker segments through the compositional index profile. An overview of CISA is illustrated in Figure 4.1. Both steps are described in the proceeding sections.

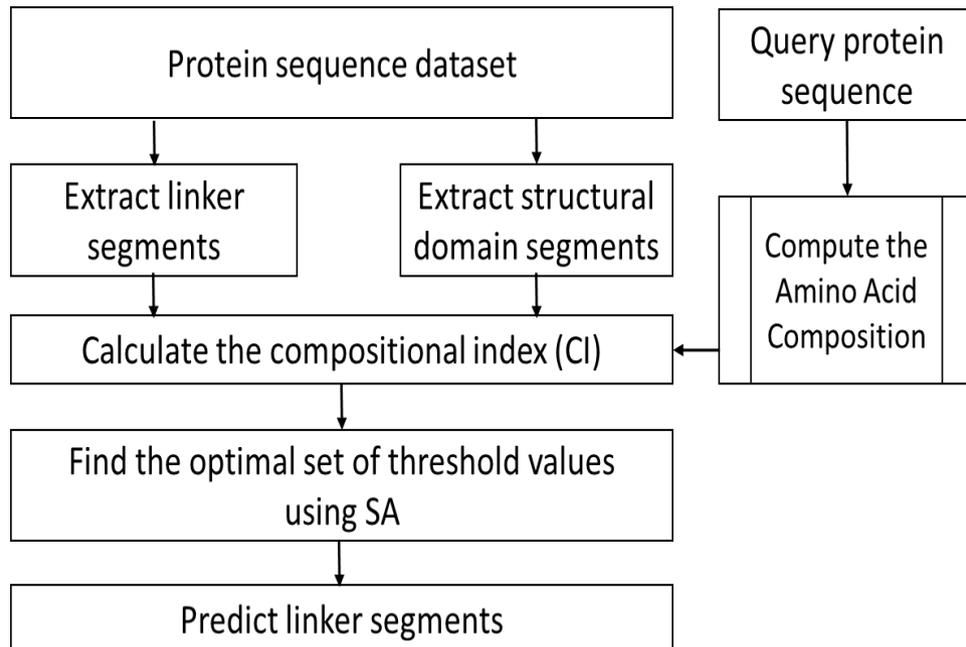


Figure 4.1: CISA overview. The method consists of two main steps; calculating the compositional index and then apply Simulated Annealing to refine the prediction by detecting the optimal set of threshold values that distinguish linkers from non-linkers.

## 4.1 Compositional Index

From each protein sequence  $s_i$  in the protein sequences database  $S^*$ , known linker segments and domain segments are extracted and saved in two datasets  $S_1$  and  $S_2$ , respectively. The compositional index  $c_i$  of the amino acid  $i$  is calculated to represent the preference of this amino acid residue to appear in linker segments:

$$c_i = -\ln\left(\frac{f_i^{linker}}{f_i^{domain}}\right) \cdot \left(\frac{k}{a_i}\right) \quad (4.1)$$

where  $f_i^{linker}$  and  $f_i^{domain}$  are the frequencies of amino acid residue  $i$  in linker and domain regions, respectively. This is inspired by DomCut method [4] which was discussed in section 2.1.1. However, the information encoded in the linker index (LI) is insufficient to precisely predict linker segments. Therefore, we used the compositional index proposed by [156] in which AA compositional knowledge was combined. The typical

AA Composition (AAC) contains 20 components, each of which reflects the normalized occurrence frequency for one of the 20 natural AAs in the query sequence. The AAC in this case is denoted by  $a_i$ . Since domain regions are usually longer than linker regions, AAC for the AA residues are more likely to appear in domains is expected to be greater than those of linkers. So multiplying LI by AAC as in [5] will scale linker regions less than domain regions. In contrast, LI is now multiplied by  $\frac{k}{a_i}$ , where  $k$  is a constant and therefore, LI of linker regions will be scaled up greater than LI of domain regions. In this case linker regions will have deeper troughs in the compositional index profile than other regions. Each residue in the query protein sequence is represented by its corresponding compositional index  $c_i$ . Subsequently, the index values are averaged over a window that slides along the length of the sequence. To calculate the average compositional index value  $m_j^w$  at position  $j$  in a protein sequence  $s$  of length  $L$  residues, using a sliding window of size  $w$ , we followed [156] and applied the following formula:

$$m_j^w = \begin{cases} \frac{\sum_{i=1}^{j+(w-1)/2} c_{si}}{j+(w-1)/2} ; & 1 \leq j \leq (w-1)/2 \\ \frac{\sum_{i=j-(w-1)/2}^{j+(w-1)/2} c_{si}}{j+(w-1)/2} ; & (w-1)/2 < j \leq L - (w-1)/2 \\ \frac{\sum_{i=j}^L c_{si}}{L-j+1+(w-1)/2} ; & L - (w-1)/2 < j \leq L \end{cases} \quad (4.2)$$

where  $L$  is the length of the protein and  $s_i$  is the amino acid at position  $i$  in protein sequence  $s$ .

Since using a fixed sliding window size could be biased towards a fixed linker region length, various odd window sizes are examined. The averaging is also carried out over this range according to the following formula:

$$\bar{m}_j = \frac{\sum_{l=0}^{(e-b)/2} m_j^{b+2l}}{((e-b)/2) + 1}, \quad j = 1, \dots, L \quad (4.3)$$

where  $b$  and  $e$  are odd averaging window sizes, and  $3 \leq b < e$ .

## 4.2 Detecting the Optimal Set of Threshold Values Using Simulated Annealing

Simulated Annealing is a simple easily-applicable optimization technique introduced by Kirkpatrick *et al.* [157] as a computational analogous to the annealing process which is the heating and controlled cooling of a metal to increase the size of its crystals and reduce their defects. The function to be optimized in SA is called the energy,  $E(x)$ , of the state  $x$ , and during that, a parameter  $T$ , the computational temperature, is lowered throughout the process. SA is an iterative trajectory descent algorithm that keeps a single candidate solution at any time [158, 159].

The major advantage of SA is its ability to avoid being trapped in local optima because the algorithm applies a random search which does not only accept changes that improve the objective function, but also some changes that temporarily worsen it [160, 161]. Geman and Geman [162] presented evidence that SA guarantees to converge to the global optimum if the cooling schedule is adequately slow. On the other hand, Salamon *et al.* [163] and Ingber [164] reported through experience that SA shows a very effective optimization performance even with relatively rapid cooling schedules [165]. The run time of SA has the complexity of  $O(n^2 \log n)$  [166].

SA is commonly found in industry and provides good optimization results [158, 159]. It has been examined and showed well performances in a variety of single-objective and multi-objective optimization applications as reported by several researchers. Some of these applications are wireless telecommunications networks [165, 159, 167], nurse scheduling problems [168], high-dimensional and complex nanophotonic engineering problems [169], pattern detection in seismograms [170], dynamic pathway identification from gene expression profiles [171], eukaryotic cell cycle regulation [172], gene network

model optimization [173], biclustering of gene expression data [174], and multiple biological sequence alignment [175, 176, 177]. However, examining SA in protein structure problems is not well addressed in the literature. Due to this reason, in addition to the previously mentioned SA features, we have decided to examine SA in domain-linker prediction.

As mentioned earlier, a dynamic threshold value is required to separate domains from linker regions. In our case, the compositional index values,  $m_j^w$ , are used in conjunction with SA algorithm. This is done by first dividing each protein sequence into chunks. Starting from a random seed  $S_o$ , which is a set of threshold values of the compositional index of these chunks, SA will attempt to simultaneously maximize both prediction recall  $R(S)$  and precision  $P(S)$ , which can be considered as a multi-objective optimization problem with both  $R(S)$  and  $P(S)$  are the fitness functions and the set of threshold values,  $S$ , is the candidate solution space, or individual representation. That is:

$$\max y = f(S) \equiv (R(S) \text{ and } P(S)) \quad (4.4)$$

Precision and recall should be maximized simultaneously. A perfect precision score can be achieved by simply assigning "domain" to all the protein sequence residues ( $FP = 0$ ), and a perfect recall score can be simply achieved by assigning "linker" to all residues ( $FN = 0$ ). However, a truly accurate predictor should assign the correct categories and only the correct categories by maximizing precision and recall at the same time, and accordingly, maximizing the F1 score.

In our case, SA will accept a transition from state  $S_1$  to another state  $S_2$  if  $S_2$  dominates  $S_1$ , that is if  $S_2$  is not worse for all objectives than  $S_1$  and wholly better for at least one objective. In other words, SA will accept a transition that leads to one

of the following three conditions: an increase in both recall and precision, an increase in recall if precision is not changed, or an increase in precision if recall is not changed.

That is:

$$\left\{ \begin{array}{l} R(S_2) > R(S_1) \text{ and } P(S_2) \geq P(S_1), \\ or \\ P(S_2) > P(S_1) \text{ and } R(S_2) \geq R(S_1) \end{array} \right. \quad (4.5)$$

SA will also accept a transition from state  $S_1$  to  $S_2$  if  $S_2$  does not dominate  $S_1$  with a probability of  $e^{(-\Delta f/T)}$ , where  $\Delta f = f(S_2) - f(S_1)$ , and  $T$  is the temperature parameter which expected to be reduced over time during the process and therefore, the possibility of accepting such transitions is decreased. The method is summarized in Algorithm 1.

---

**Algorithm 1** Inter-Domain Linker Prediction Optimization
 

---

Set  $S_0$  as an initial candidate solution:

    Divide the protein sequence into chunks

    Assign a random initial threshold of each chunk

Calculate CI

Classify each AA as linker (1) or domain (0) according to its CI value with respect to the corresponding chunk threshold

Calculate the fitness functions  $R(S_0)$  and  $P(S_0)$

$T_0 \leftarrow$  *Initial temperature*

$\alpha \leftarrow$  *Temperature decay*

Maximize the fitness functions:

**for**  $n = 1$  to *Number of Chunks* **do**

$T \leftarrow$  *Temperature*

**repeat**

        Make a transition Tr:

            randomly increase or decrease threshold of  $n$

$S \leftarrow Tr(S_0)$

        Classify each AA as linker or domain

        Calculate  $R(s)$  and  $P(s)$

$\Delta R \leftarrow R(S) - R(S_0)$  and  $\Delta P \leftarrow P(S) - P(S_0)$

**if** ( $\Delta R > 0$  and  $\Delta P \geq 0$ ) or ( $\Delta P > 0$  and  $\Delta R \geq 0$ ) **then**

            accept transition

**else if**  $random[0, 1) < exp(-\frac{\Delta R + \Delta P}{T})$  **then**

            accept transition

**end if**

$T \leftarrow \alpha \times T$

**until** stopping criteria is met

**end for**

**return**  $S$  as the set of optimal threshold values for the protein sequence chunks

**return**  $R(S)$  and  $P(S)$  as the final recall and precision, respectively

---

### 4.3 Experimental Results and Discussion

To illustrate the improvement of our modified compositional index over both the linker index of [4] and the compositional index of [156, 5], three profiles of a protein 1au7\_A are plotted as shown in Figure 4.2. The 1au7\_A protein sequence of Chain A, Pit-1 MutantDNA Complex has 146 AA residues and contains an actual domain linker located in the positions from 74 to 109 as retrieved from the National Center for Biotechnology Information (NCBI)<sup>1</sup> and indicated by the horizontal arrow in the figure. The figure shows that the modified compositional index can separate linker regions from domain regions more accurately and sharply than those of [4] and [5]. Figure 4.2(c) shows how the trough in the linker region is deeper than those of Figure 4.2(a) and (b), respectively. We can also notice that the profile in Figure 4.2(b) has a second trough indicating a false linker in the right side of the profile which is deeper than the actual linker's trough.

Another example is illustrated in Figure 4.3 based on the 1f6f\_C protein which has 210 AA residues and one linker as retrieved from NCBI and indicated by the horizontal arrow. Figure 4.3(a) (the linker index of [4]) and 4.3(b) (the compositional index of [5]) show more than one trough indicating false linkers and the index values of these false linkers are less than those of the actual linker. However, Figure 4.3(c) clearly shows that, according to our proposed modified formula, the residues in the actual linker regions have lower index values than those of other residues which allows to easily find a separation threshold.

As shown in Figures 4.2 and 4.3, having a static threshold cannot precisely separates linkers from domain regions, and therefore, a dynamic threshold is required.

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/>

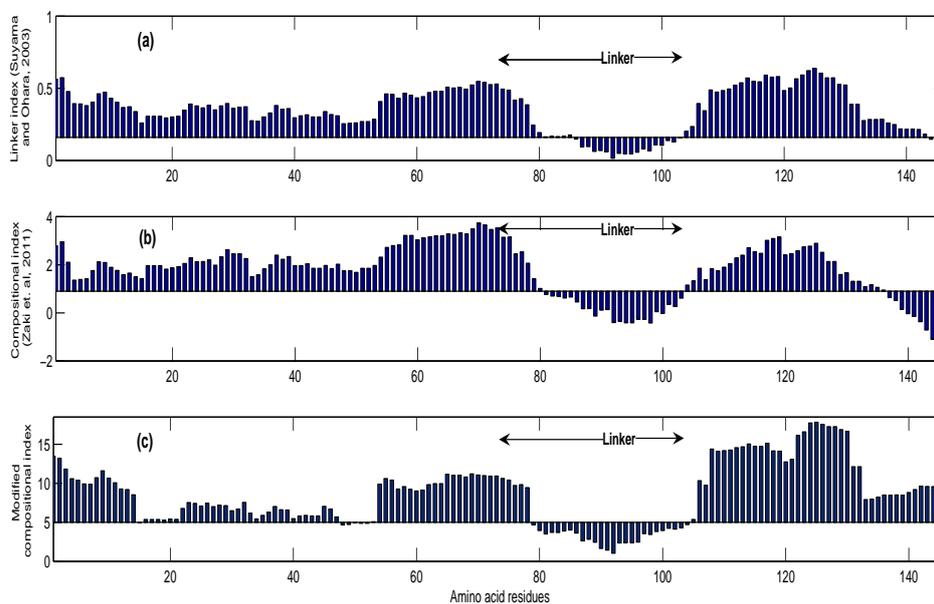


Figure 4.2: Comparison between (a) linker index of [4], (b) compositional index of [5], and (c) the modified compositional index profiles for 1au7\_A protein. The horizontal arrow indicates the actual linker region located in the positions from 74 to 109 (according to NCBI). Each AA below a defined threshold is considered to be a linker. Although all three methods predicted the linker in this case, the linker predicted by CISA is more accurate when compared to the actual linker.

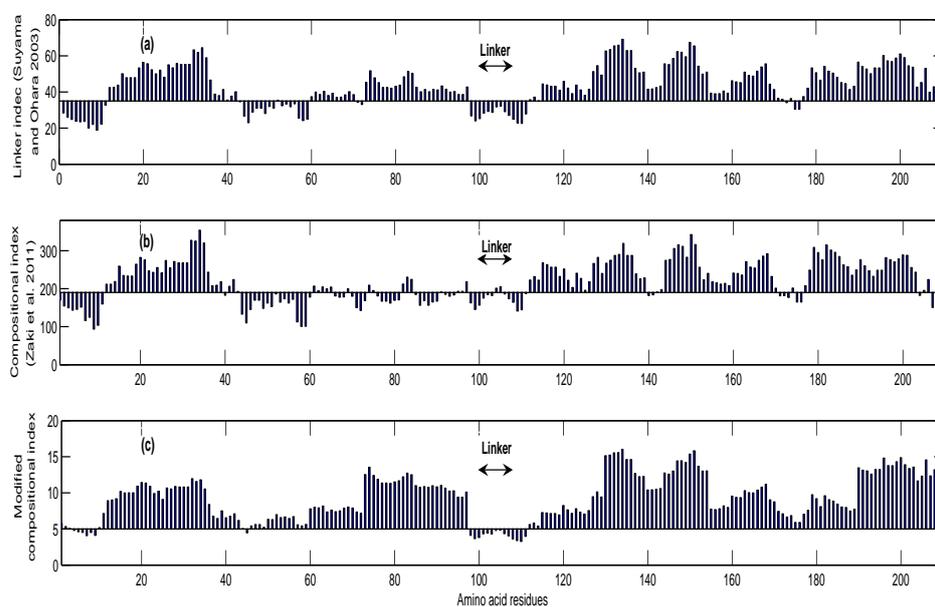


Figure 4.3: Comparison between (a) linker index of [4], (b) compositional index of [5], and (c) the modified compositional index profiles for 1f6f\_C protein. The horizontal arrow indicates the actual linker region (according to NCBI). Each AA below a defined threshold is considered to be a linker. Although all three methods predicted the linker in this case, the linker predicted by CISA is more accurate when compared to the actual linker. Both linker index of [4] and compositional index of [5] have predicted two clear false linkers (regions  $\approx$  1-10 and 44-60, respectively.)

We applied the SA technique to detect the optimal set of threshold values that will separates linkers from domain regions along the protein sequence.

We evaluated the performance of CISA using DomCut/Swiss-Prot protein dataset which was prepared by [4] using one-against-all cross validation and explored different chunk sizes  $\{5, 10, 18, 36\}$  where 36 is the average linker size within the dataset. CISA was able to achieve an average recall of 0.89, precision of 0.80 and F1-measure of 0.84 on a window size of 25 residues and a chunk of 5 residues Figure 4.4 presents these evaluation metrics at different chunk sizes.

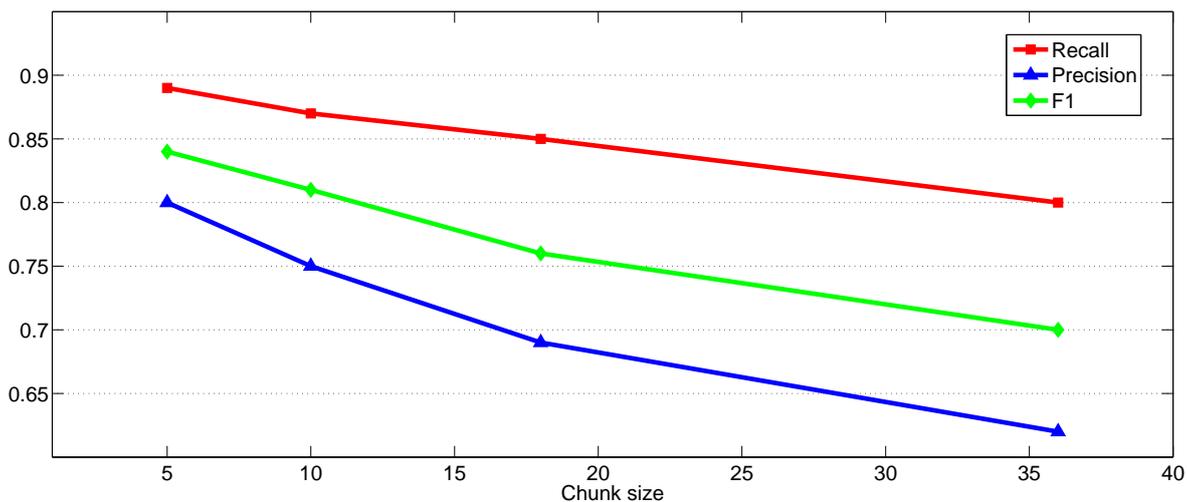


Figure 4.4: Recall, precision, and F1-measure at a window size of 25 and at different chunk sizes (5 to 36) using DomCut/Swiss-Prot dataset.

In the second experiment, we evaluated the performance of our method on 151 protein sequences of DS-All dataset including 182 linker and 332 domains. In this experiment DomCut dataset was used to generate the linker index of each AA before using them to predict the domain-linker regions in DS-All dataset. Several odd sliding window sizes  $w$  in the range of 5 to 25 AAs are explored for computing the compositional index  $m_j^w$  according to equation 4.2. It was noticed that the best results were achieved when  $w \geq 19$  as shown in Figure 4.5. Further, we tested the averaging  $\bar{m}_j$  over a range

of 5 to 25 AAs according to equation 4.3. This process takes a longer computational time without a significant improvement in the prediction accuracy as shown in Figure 4.5. As a result, we decided to set  $w$  to 25 in all of our experimental works. To optimize the scaling constant  $k$ , we examined three values  $\{1, 10, 100\}$ . Based on Equation 4.1, we found that F1-score is slightly higher when  $k = 100$  than  $k = 10$ , and significantly higher than that at  $k = 1$ .

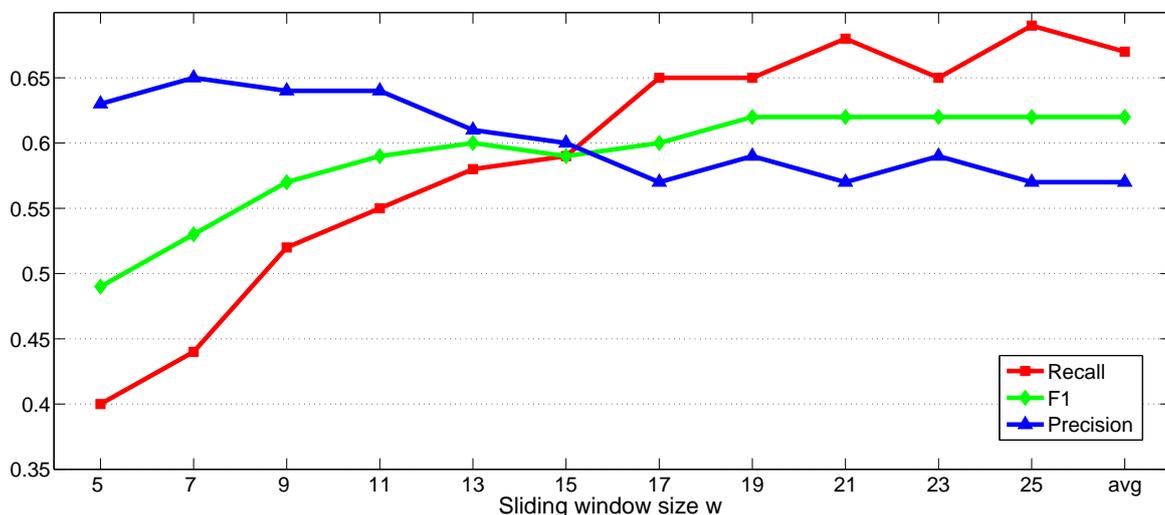


Figure 4.5: Recall, precision and F1-measure based on DS-All dataset by [6] and [7]. The sliding window sizes  $w$  is set in the range of 5 to 25 AAs. The average value of the sliding window sizes (*avg*) is also included.

We have also explored several chunk sizes  $\{5, 10, 13\}$ , where 13 is the average linker size among the dataset. Figure 4.6 presents these evaluation metrics at different chunk sizes. We were able to achieve an average prediction recall of 0.78, precision of 0.79 and F1-measure of 0.79 when the chunk size was set to 5 AA long.

Although our algorithm selects a random chunk in the initial iteration, it can be easily modified to scan the protein sequence from left to right in order to cover the whole chunks across the chain. One of the challenges that we faced during the evaluation step of the algorithm is the division by zero during the calculation of the precision. This is normally happens at the early stages where no AA regions are predicted as linkers

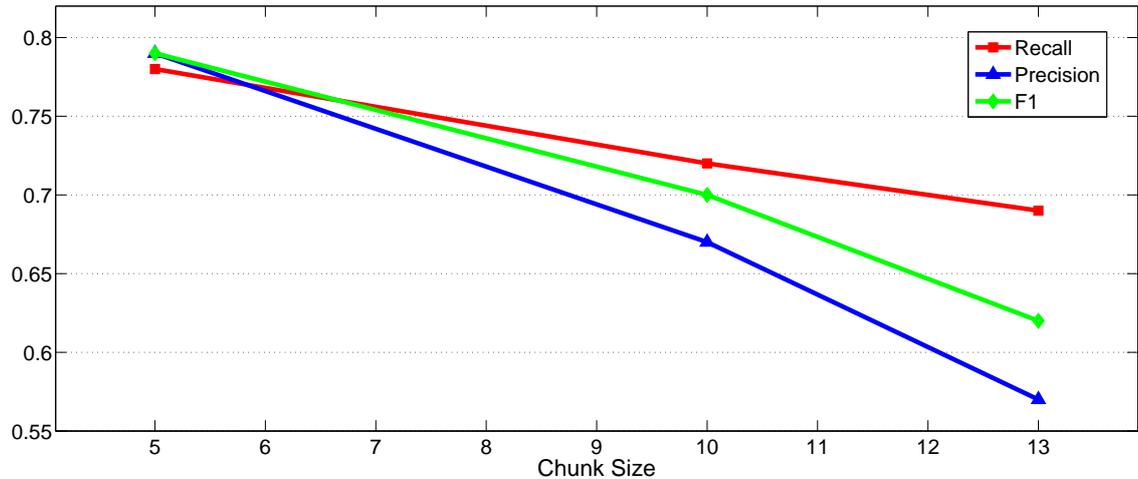


Figure 4.6: Recall, precision and F1-measure at a window size of 25 and at different chunk sizes based on DS-All dataset.

and, therefore, the true positive (TP) and false positive (FP) are zeros. To overcome this challenge, we designed the algorithm in a way to reject such state and immediately performs a new transition.

Another challenge is the fact that the recall  $R(S)$  and precision  $P(S)$  are not continuous functions. In other words, a change in  $S$  (the set of threshold values) may cause a jump in the values of  $R(S)$  and  $P(S)$ , or it may cause no change in both values. At the same time the transition,  $\Delta S$  should be maintained, which is a change in a threshold of one chunk, within a reasonable range that we set to be  $\frac{1}{10}$  of the compositional index range. Therefore, the algorithm should perform several transitions till it passes from state  $S_1$  to a more dominant state  $S_2$ . However, while performing, these transitions,  $\Delta R$  and  $\Delta P$  will be zeros while the algorithm has not yet converged to the global maximum. Therefore, we did not consider having  $\Delta R = 0$  and  $\Delta P = 0$  as a stopping criteria. Instead, we set the number of iterations to 20 per chunk.

One of the SA algorithm issues we had to deal with is the random seed, or initialization issue. Depending on the initial state, SA performs differently and returns different outputs. This issue can be addressed by setting a predefined initial threshold value for the whole input sequence residues. We set this initial threshold to be the

average value of the CI as this average value is somehow in the middle of the CI profile which can help SA to converge more efficiently by either stepping-up the threshold in linker segments or stepping-down the threshold in domain segments.

### 4.3.1 Performance Comparison

Based on the DS-All dataset, the performance of CISA was compared to the currently available domain linker prediction approaches as shown in Figure 4.7. CISA was able to outperform 6 of the state-of-the-art domain-linker prediction approaches in terms of recall, precision and F1-score. As shown in Table 4.1, the performance of CISA was also compared to the recent predictor developed by [178] and DomCut based on the Swiss-Prot/DomCut dataset. CISA was also able to show considerable improvement in prediction accuracy.

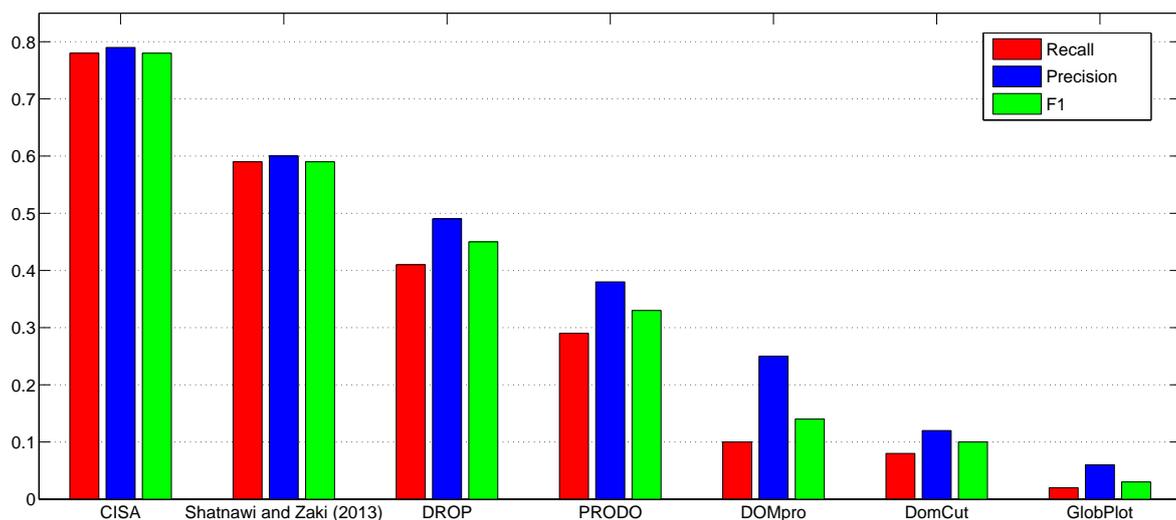


Figure 4.7: CISA performance compared to the state-of-the-art predictors based on the DS-All dataset.

### 4.3.2 Biological Relevance

To demonstrate the performance of CISA, Figure 4.8(a) shows the compositional index profile for 1au7\_A protein sequence in DS-All dataset which contains 146

<b>Method</b>	<b>Recall</b>	<b>Precision</b>	<b>F1</b>
CISA	0.89	0.80	0.84
Shatnawi and Zaki[178]	0.56	0.84	0.67
DomCut	0.54	0.50	0.52

Table 4.1: CISA performance comparison using Swiss-Prot/DomCut dataset.

AA residues and has two domains and a domain-linker in the region from 74 to 109. The figure also shows the optimal threshold values achieved by CISA. It is shown that the compositional index threshold values at linker segments are raised by the algorithm while threshold values of domains are reduced. In this case the compositional index values of a linker region will be lower than its associated threshold values while the compositional index values of a domain region will be higher than its associated thresholds, and this, in turn, improve the prediction. The three dimensional structure of this protein is shown in Figure. 4.8(b) which shows the two domains in red and green retrieved from NCBI<sup>2</sup>.

Identification of domain linkers locations is often the first step in protein folding and function annotations. Another example that illustrates how CISA can furthermore assists in detecting important domains by identifying linkers is the detection of three important conservative domains in the breast cancer type 1 (BRCA1) susceptibility protein isoform 4 [Homo sapiens] which consists of 759 AAs. Figure 4.9 presents the compositional index profile for this protein and the threshold values achieved by CISA. It is shown that the proposed algorithm can accurately detect the domain linkers which leads to the identification of three important domains. The first domain is RING-finger domain which is a specialized type of Zn-finger that binds two atoms of zinc, involved in mediating protein-protein interactions, and identified in proteins with a wide range of functions such as viral replication, signal transduction, and development. This domain

---

<sup>2</sup><http://www.ncbi.nlm.nih.gov/>

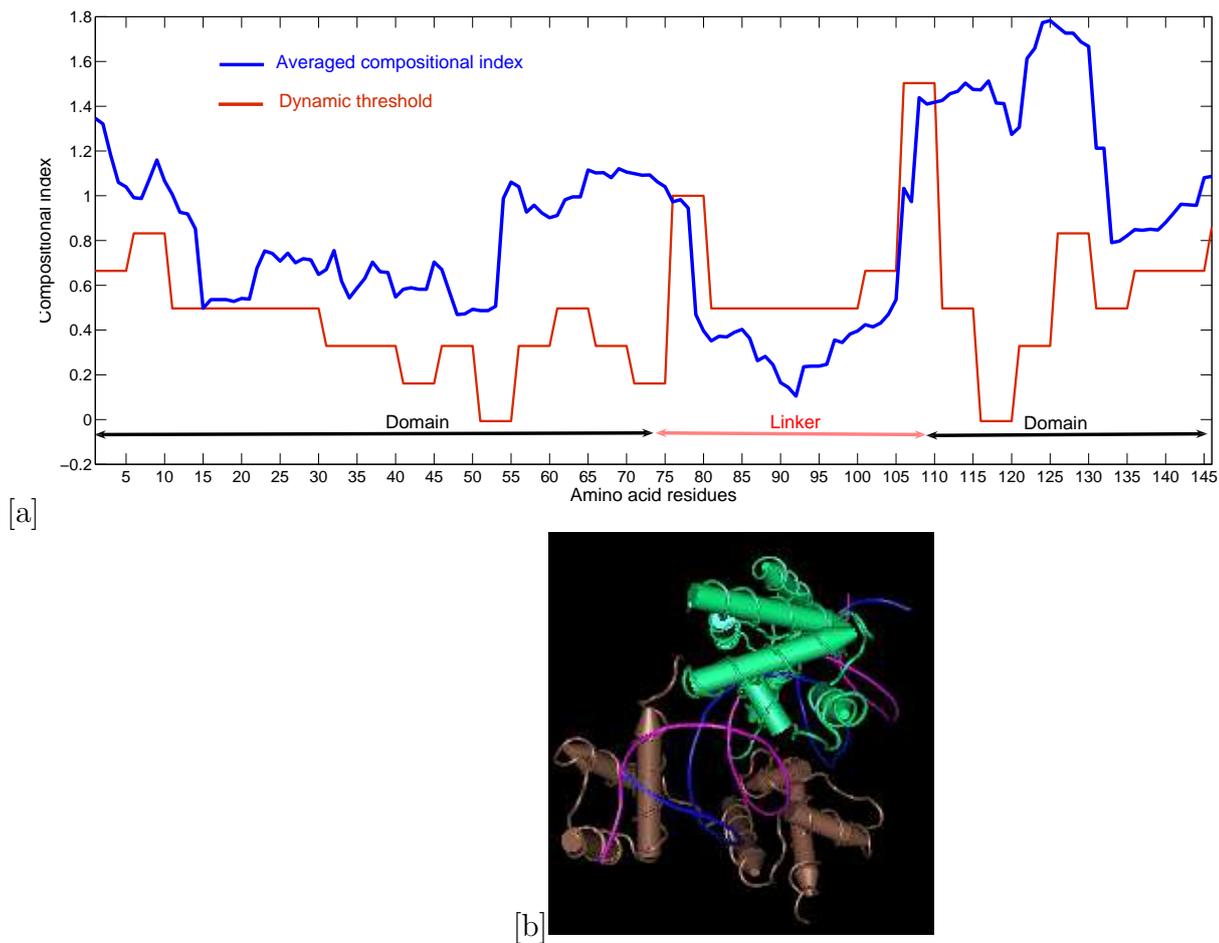


Figure 4.8: Protein 1au7\_A in DS-All dataset which has 146 AA residues containing two domains. (a) The compositional index (CI) profile (blue) and the optimal threshold values returned by the algorithm (red). The two domains and the linker are shown at the bottom as labeled in the DS-All dataset. It is shown how the algorithm returns threshold values higher than CI in the linker region while it returns threshold values lower than CI in domain regions. (b) The 3D structure for this protein showing the two domains.

is located at positions 23 to 68. The other two domains are Breast Cancer Suppressor Protein (BRCA1) carboxy-terminal domains. They are found within many DNA damage repair and cell cycle checkpoint proteins. These two domain are located in positions from 546 to 620 and from 659 to 738, respectively.

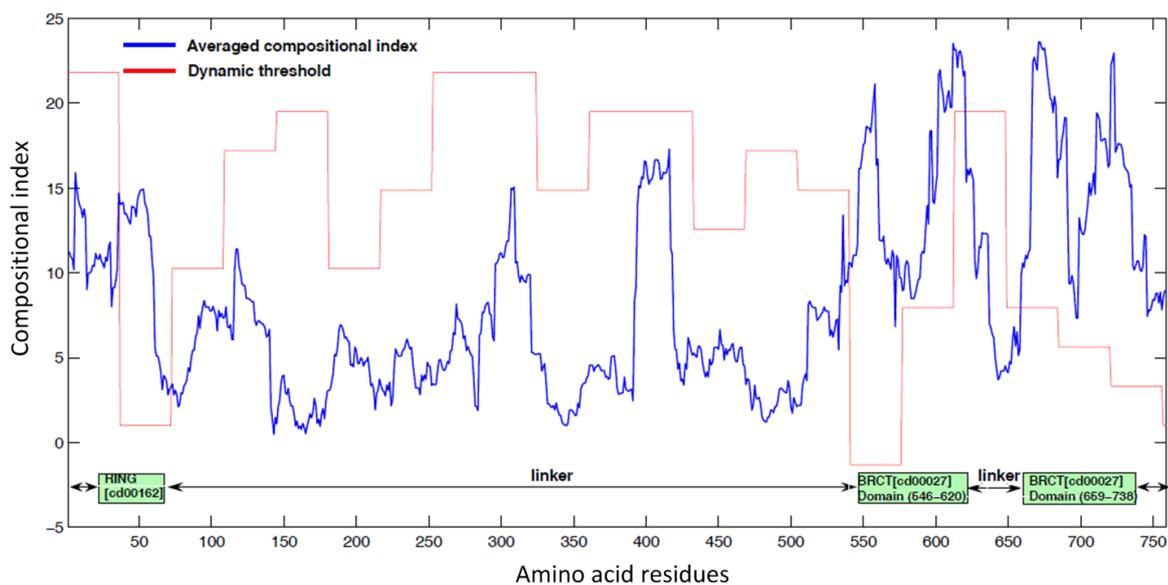


Figure 4.9: The CI profile based on the Breast cancer type 1 susceptibility protein is shown in blue and the optimal threshold values achieved by CISA are shown in red. The three domains according to the NCBI's conserved domain database are represented by the green boxes. It is shown that CISA was able to discriminate between linker regions (threshold  $>$  CI) and domain regions (threshold  $<$  CI).

# Chapter 5: Random Forest Approach for Domain and Linker Prediction

In this chapter, we present our second contribution which is identifying structural domains based on linker knowledge. To include biological knowledge to the compositional index which was introduced in Chapter 4, we combine the compositional index with several AA physiochemical properties to construct a novel protein profile. This profile is then used to build a machine learning classifier to predict novel domains and linkers. We utilize a nature-inspired machine-learning model called Random Forest. Section 5.1 describes the feature extraction stage while Section 5.2 describes the Random Forest model. Experimental results are presented and discussed in Section 5.3.

## 5.1 Feature Extraction

To extract AA features from a protein, a sliding window technique is used. For each sequence in the protein dataset, we slide an averaging window across the sequence from the *N*-terminal to the *C*-terminal. A number of important features of a protein, located within the sliding window, are extracted. These features are the compositional index which was introduced in Section 4.1, AA hydrophobicity, and other AA physiochemical properties including side-chain charge, side-chain polarity, aromaticity, size, and electronic properties.

### 5.1.1 Hydrophobicity Profile

Hydrophobicity is a physical property of a substance to repel water and it is a major factor in protein stability. The hydrophobic effect plays a key role in the spontaneous folding of proteins. It can be defined as the free energy required to transfer amino-acid side-chains from cyclohexane to water [179]. Table 5.1 illustrates hydrophobicity index in kilo-calories per mole for each of the twenty AAs of proteins at a pH of 7. Several researchers selected hydrophobicity as the main feature among many other properties in protein structure prediction [179, 180, 181, 182].

Amino acid	Hydrophobicity index	Amino acid	Hydrophobicity index
I	4.92	Y	-0.14
L	4.92	T	-2.57
V	4.04	S	-3.40
P	4.04	H	-4.66
F	2.98	Q	-5.54
M	2.35	K	5.55
W	2.33	N	-6.64
A	1.81	E	-6.81
C	1.28	D	-8.72
G	0.94	R	-14.92

Table 5.1: Hydrophobicity index (kcal/mol) of amino acids in a distribution from non-polar to polar at pH=7 [182].

In literature, various hydrophobicity scales have been thoroughly examined for protein sequence classification and prediction tasks. David [183] concluded that the Rose scale [184] was superior to all others when used for protein structure prediction. The Rose scale in Table 5.2 is correlated to the average area of buried AAs in globular proteins. However, Korenberg *et al.* [181] pointed out several key drawbacks with Rose scale. Since it is not a one-to-one mapping, different amino-acid sequences can have identical hydrophobicity profiles; the scale covers a narrow range of values while causing

Amino acid	Hydrophobicity index	Amino acid	Hydrophobicity index
A	0.74	L	0.85
R	0.64	K	0.52
N	0.63	M	0.85
D	0.62	F	0.88
C	0.91	P	0.64
Q	0.62	S	0.66
E	0.62	T	0.70
G	0.72	W	0.85
H	0.78	Y	0.76
I	0.88	V	0.86

Table 5.2: Rose hydrophobicity scale. The scale is correlated to the average area of buried AAs in globular proteins [182].

Amino acid	Hydrophobicity index	Amino acid	Hydrophobicity index
C	1,1,0,0,0	G	0,0,0,-1,-1
F	1,0,1,0,0	T	0,0,-1,0,-1
I	1,0,0,1,0	S	0,0,-1,-1,0
V	1,0,0,0,1	R	0,-1,0,0,-1
L	0,1,1,0,0	P	0,-1,0,-1,0
W	0,1,0,1,0	N	0,-1,-1,0,0
M	0,1,0,0,1	D	-1,0,0,0,-1
H	0,0,1,1,0	Q	-1,0,0,-1,0
Y	0,0,1,0,1	E	-1,0,-1,0,0
A	0,0,0,1,1	K	-1,-1,0,0,0

Table 5.3: SARAHI hydrophobicity scale. Each AA is assigned a five-bit code in descending order of the binary value of the corresponding code where the right-half is the negative mirror image of the left-half. The 10 most hydrophobic residues are positive, and the 10 least hydrophobic residues are negative [182].

some AAs to be weighted more heavily than others. To overcome this problems, the SARAHI scale was introduced [181]. SARAHI assigns to each AA a unique five-bit signed code, where exactly two bits are non-zero, as illustrated in Table 5.3 where the right-half is the negative mirror image of the left-half. The ten most hydrophobic residues are positive while the ten least hydrophobic residues are negative.

In this work, we experimentally tested the three above mentioned hydrophobicity scales where SARAHI scale showed a slightly better prediction accuracy. Thus, we used SARAHI in the construction of our AA feature set.

### 5.1.2 Physiochemical Properties

In addition to hydrophobicity, we considered several physiochemical properties of AAs as features including electric charge, polarity, aromaticity, size, and electronic property. AAs are categorized according to each physiochemical property as in Table 5.4 [185, 186, 187]. Each physiochemical property of an AA is based on its side-chain propensity and has its own characteristics. Physiochemical properties play important role in recognizing the behavior of the AAs and its interactions with other AAs. These interactions have significant impact on the formation, folding, and stabilization of protein 3D structures. For example, polar and charged AAs are able to form hydrogen bonds, and thus, they cover the molecules surfaces and are in contact with solvents. Positively and negatively charged amino acids form salt bridges. Polar amino acids are hydrophilic, whereas non-polar amino acids are hydrophobic, which are used to twist protein into useful shapes [188, 189].

Property	Value	Amino acids
Charge	Positive	H, K, R
	Negative	D, E
	Neutral	A, C, F, G, I, L, M, N, P, Q, S, T, V, W, Y
Polarity	Polar	C, D, E, H, K, N, Q, R, S, T, Y
	Non-polar	A, F, G, I, L, M, P, V, W
Aliphatic/Aromatic	Aliphatic	I, L, V
	Aromatic	F, H, W, Y
	Neutral	A, C, D, E, G, K, M, N, P, Q, R, S, T
Size	Small	A, G, P, S
	Medium	D, N, T
	Large	C, E, F, H, I, K, L, M, Q, R, V, W, Y,
Electronic	Strong donor	A, D, E, P
	Weak donor	I, L, V
	Neutral	C, G, H, S, W
	Weak acceptor	F, M, Q, T, Y
	Strong acceptor	K, N, R

Table 5.4: Amino acid classification according to their physiochemical properties [185, 186, 187].

### 5.1.3 Protein Sequence Representation

Each sequence in the dataset is replaced by its corresponding properties; compositional index, hydrophobicity, charge, polarity, aromaticity, size, and electronic property. These values are then averaged over a window that slides along the length of each protein sequence starting from the *N*-terminal towards the *C*-terminal. To calculate the average feature values  $X_j^w$  at position  $j$  in a protein sequence  $S$ , using a sliding window of size  $w$ , we map feature values into numbers and then apply the following formula:



## 5.2 Random Forest Model

Random Forest (RF) [136] is an ensemble learner that constructs a multitude of decision trees with randomly selected features during training time and outputs the class that is the mode of the classes output by individual trees. Each decision tree grows as follows: for a training set of  $N$  cases and  $M$  variables, sample  $n$  cases with replacement from the original data to grow the tree. A number  $m \ll M$  is specified such that at each node  $m$  variables are selected randomly to best split the nodes. Each tree grows as large as possible. The error of RF depends on the strength of each individual tree and the correlation between them [190]. RF algorithm is depicted in Figure 5.2.

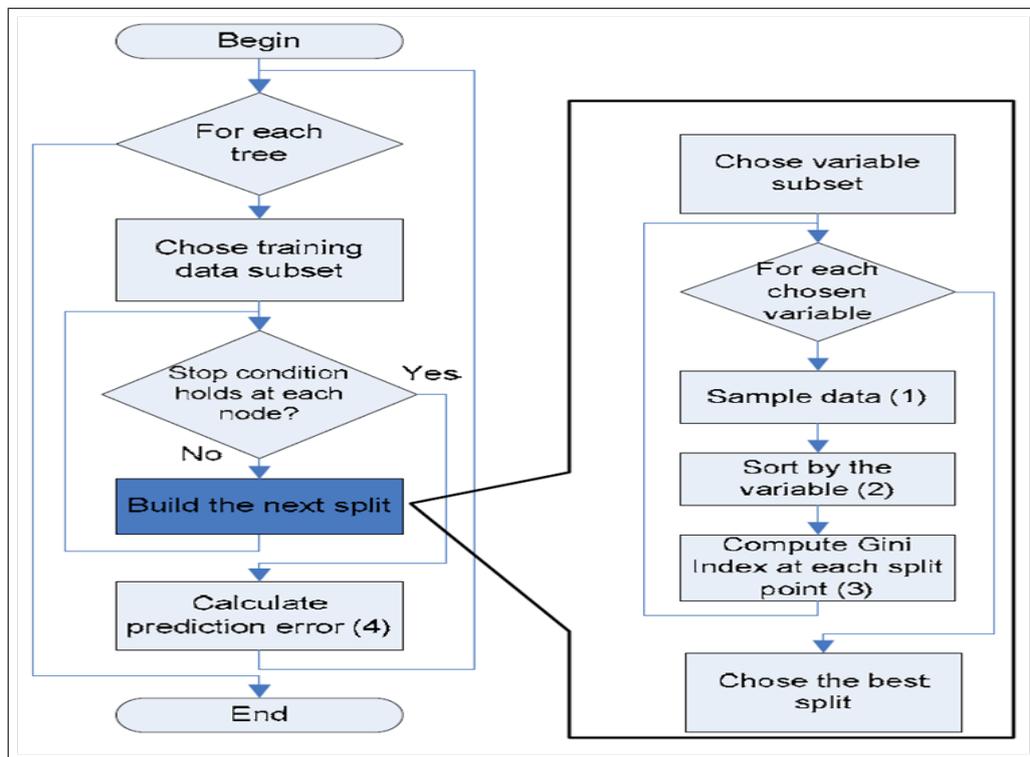


Figure 5.2: Random Forest Algorithm

Due to its averaging strategy, RF classifier is robust to outliers and noise, avoids overfitting, is relatively fast, simple, easily parallelized, and performs well in many classification problems [136, 137]. RF shows a significant performance improvement over

the single tree classifiers such as CART and C4.5. RF model interprets the importance of the features using measures such as decrease mean accuracy or *Gini* importance [138]. RF benefit from the randomization of decision trees as they have low-bias and high variance. RF has few parameters to tune and less dependent on tuning parameters [139, 140].

Ensemble methods including RF, bagging, and boosting have been increasingly applied to bioinformatics. When compared to bagging and boosting ensemble methods, RF has a unique advantage of using multiple feature subsets which is well suited for high-dimensional data as demonstrated by several bioinformatics studies [191]. Lee *et al.* [192] compared the ensemble of bagging, boosting and RF using the same experimental settings and found that RF is the most successful one. The experimental results through ten microarray datasets in [193] reported that RF is able to preserve predictive accuracy while yielding smaller gene sets compared to diagonal linear discriminant analysis, kNN, SVM, shrunken centroids (SC), and kNN with feature selection. Other advantages of RF such as robustness to noise, lack of dependence upon tuning parameters, and the computation speed have been verified by [139] in classifying SELDI-TOF proteomic data. Wu *et al.* [194] compared the ensemble methods of bagging, boosting, and RF to individual classifiers of LDA, quadratic discriminant analysis, kNN, and SVM for MALDI-TOF (matrix assisted laser desorption/ionization with time-of-flight) data classification and reported that among all methods RF gives the lowest error rate with the smallest variance. RF also has better generalization ability than Adaboost ensembles [195].

Recently, RF has been successfully employed to a wide range of bioinformatics problems including protein-protein binding sites [196], protein-protein interaction [89, 197], protein disordered regions [198], transmembrane helix [189], residue-residue

contact and helix-helix interaction [190], and solvent accessible surface area of TM helix residues in membrane proteins [199].

In our case, the feature vector constructed in the last section is used to train the RF classifier. At each node of every tree, a number of features are randomly selected and the feature which can better split the dataset is chosen among them. We set the number of selected features at each node for building the trees,  $m$ , to  $(\log_2(\textit{number of attributes}) + 1)$  as recommended by [136]. During testing, each test point is simultaneously pushed through all trees until it reaches the corresponding leave which can be either domain or linker and, in turn, RF chooses the classification with the most votes from all the trees.

### 5.3 Experimental Results and Discussion

Each AA residue in every protein sequence is represented by its corresponding feature values. These features are the compositional index that was introduced in Section 4.1, AA hydrophobicity, and other AA physiochemical properties including side-chain charge, side-chain polarity, aromaticity, size, and electronic properties. These values are then averaged over a window that slides along the length of each protein sequence according to Equation 5.1.

To find the optimal averaging window size, we tested odd window sizes in the range of 7 to 45 residues at randomly selected 50 protein sequences from DS-All dataset [31] and another randomly selected 50 protein sequences from DomCut dataset [4], and then compared the prediction performance at these windows in terms of recall, precision, and F1-score. Figure 5.3 depicts the performance measures at different sliding windows when applied to the 50 protein sequences of DS-All dataset. Figure 5.4 shows these prediction measures at different sliding windows when applied to the 50 protein sequences from DomCut dataset. As seen in these two figures, the window size of 41 showed the highest recall, precision and F-measure on both datasets. We thus set the averaging window size to 41 to obtain the final experimental results.

We set the number of selected features at each node for building the trees,  $m$ , to  $(\log_2(\text{number of attributes}) + 1)$  as recommended by [136]. We examined several values for the number of generated decision trees,  $N_{trees}$ , in the range of 10 and 500 and found that the prediction accuracy increases as  $N_{trees}$  increases as shown in Figure 5.5. However, the improvement in prediction when  $N_{trees}$  exceeds 200 is not considerable when compared with the increase in computational time and memory. Therefore, we set  $N_{trees}$  to 200 in all the conducted experiments. This also agrees with recent empirical

studies [200, 201] which reported that ensembles of size less or equal to 100 are too small for approximating the infinite ensemble prediction.

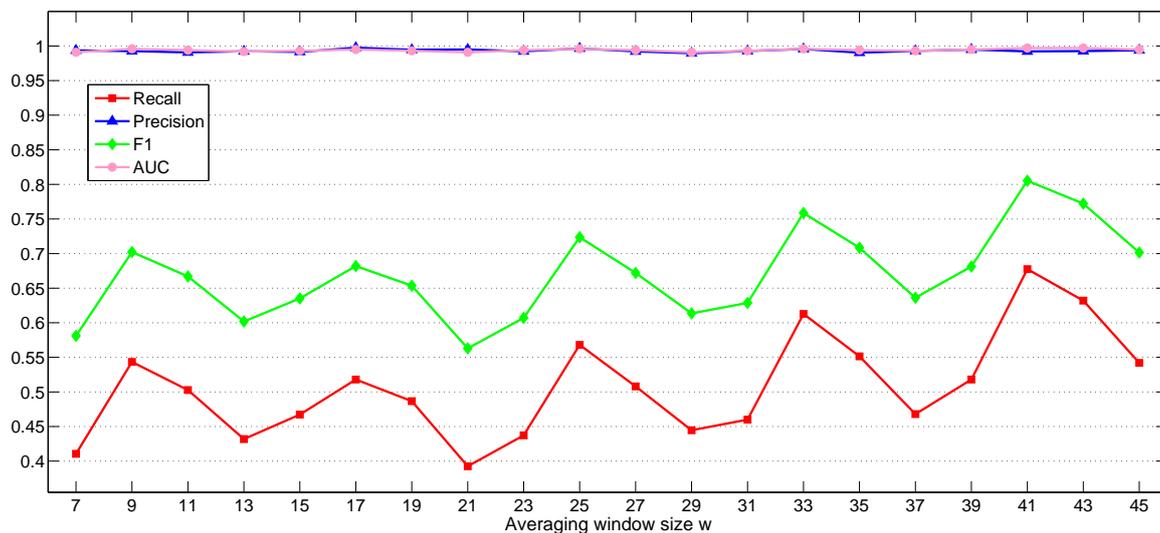


Figure 5.3: Recall, precision, F-measure, and AUC of random forest classifier at different averaging window sizes with fifty protein sequences from DS-All dataset.

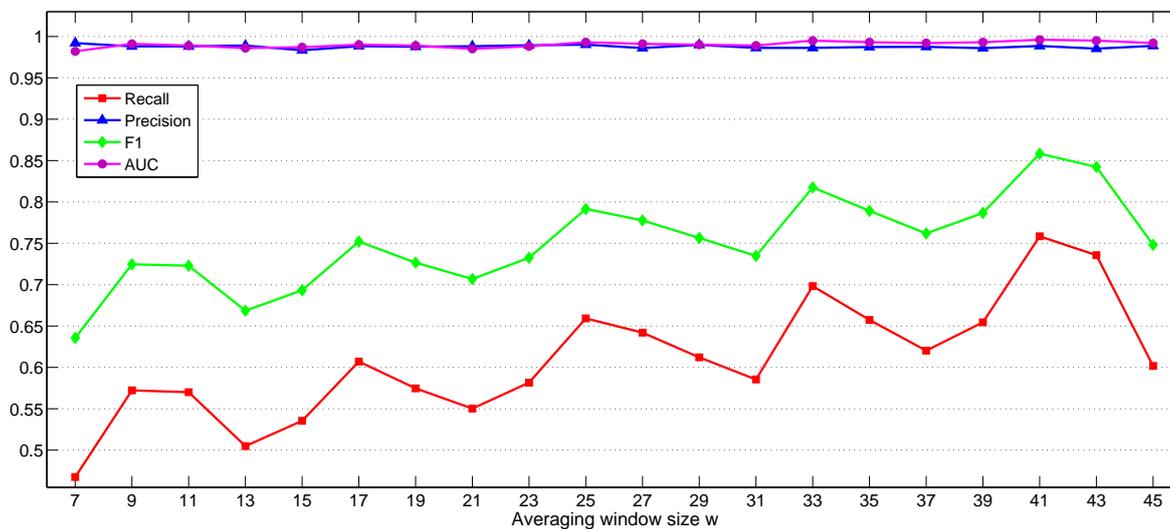


Figure 5.4: Recall, precision, F-measure, and AUC of random forest classifier at different averaging window sizes with fifty protein sequences from DomCut dataset.

The experimental results showed that the proposed approach is useful for the domain and linker identification of highly imbalanced single-domain and multi-domain proteins. Clearly, there are several advantages of the proposed approach. First, there

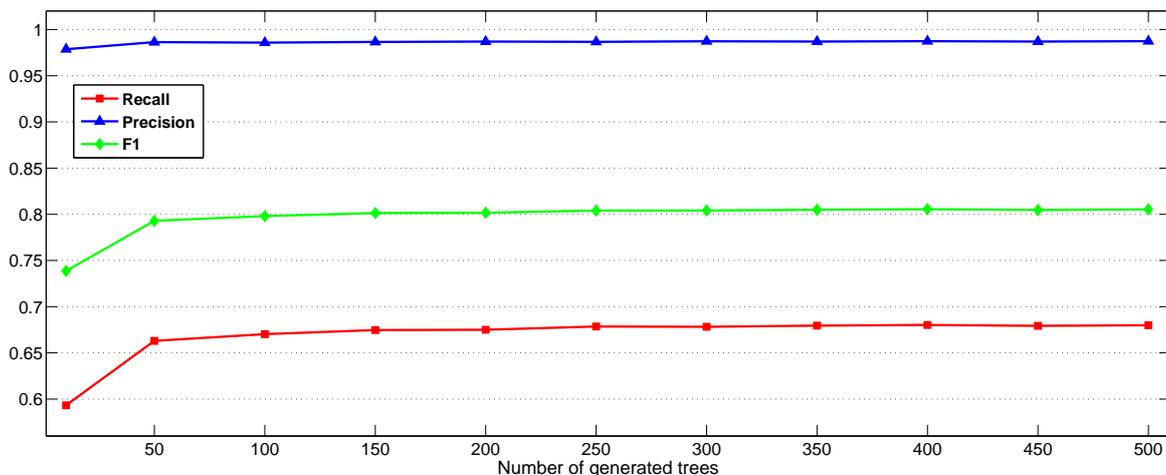


Figure 5.5: Number of generated trees optimization. Recall, precision, and F-measure at different number of generated trees performed on DS-All dataset.

are only few RF parameters that need to be tuned. Second, the better predictive performance of the proposed approach was achieved on the imbalance domain-linkers without applying any class weights or data re-sampling techniques. In other words, the proposed approach is not biased towards the majority class like most other ML models. To compare RF performance to SVM and ANN classifiers, we trained a SVM and ANN classifiers with the same protein data and found that both classified the whole protein sequences as domains. This can be explained by the fact that the training of such methods is based on adjusting the model parameters that maximize the classification accuracy (by minimizing the error rate) which is not a successful strategy in case of highly imbalanced data. Third, physiochemical properties that are used in this approach play important roles in forming the behavior of amino acids and their interactions with other amino acids and these interactions have significant impact on the formation, folding, and stabilization of protein 3D structures. Therefore, these properties are important features to distinguish structural domains from linkers. Fourth, the primary structure features that are used in this approach can be extracted with a low computational cost when compared to extracting other features such as PSSM and

protein secondary structure that are used in most of the current approaches. Generating PSSM and predicting secondary structure features are computationally expensive and time consuming. Moreover, protein secondary structures are normally predicted by SSpro [29] which reaches an accuracy of 80% only, so the incorrectly predicted secondary structures may lead to model misclassification.

To study the importance of features by finding which features contribute most to the prediction, we perform a feature selection procedure as follows. First, we measure the Information Gain (IG) of each feature and order the features according to their IG. Then, we remove the features one by one starting with the one that has least IG and find its effect on the prediction and present the results in Table 5.5. It is found that AA compositional index and hydrophobicity contribute the most while AA polarity and electric charge contributes less than other features.

Features Removed	Recall	Precision	F1
None	0.675	0.987	0.802
Polarity	0.673	0.984	0.799
Charge and Polarity	0.645	0.983	0.779
Size and all the above	0.602	0.980	0.746
Electronic and all the above	0.455	0.968	0.619
Aromaticity and all the above	0.325	0.916	0.480
Hydrophobicity and all the above	0.169	0.204	0.185

Table 5.5: Prediction measures after removing features that have less information gain using DS-All dataset.

### 5.3.1 Performance Comparison

Based on the DS-All dataset, with 10-fold cross validation, we achieved the average prediction recall of 0.68, precision of 0.99, and F-measure of 0.80. The comparisons of our approach with existing domain and linker prediction approaches [31] on DS-All dataset are summarized in Figure 5.6. Clearly, the proposed approach outperformed

the existing predictors in terms of recall, precision, and F-measure.

To prove the usefulness of our approach, it was again tested on DomCut/Swiss-Prot protein sequence dataset. Our approach again outperformed Shatnawi and Zaki’s predictor [178] as well as DomCut [4] with average recall of 0.65, a precision of 0.98, and an F-measure of 0.78 as shown in Table 5.6.

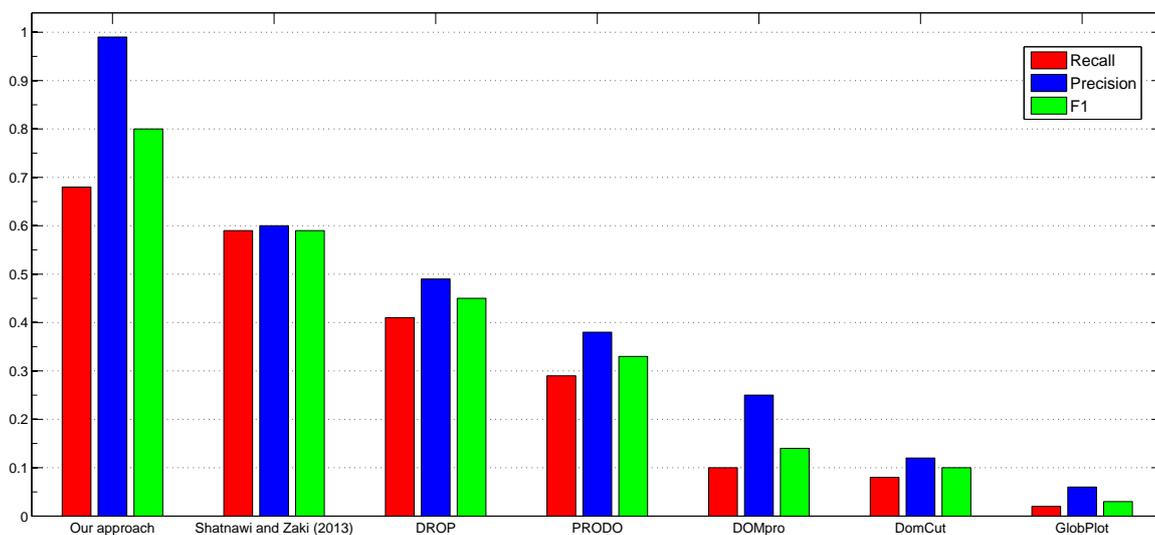


Figure 5.6: Recall, precision, and F-measure of six currently available domain boundary/linker predictors compared to our approach performed with DS-All dataset.

<b>Approach</b>	<b>Recall</b>	<b>Precision</b>	<b>F1</b>
Our Approach	0.71	0.98	0.82
Shatnawi and Zaki (2013)	0.56	0.84	0.67
DomCut	0.54	0.50	0.52

Table 5.6: Recall, precision, and F-measure using Swiss-Prot/DomCut dataset

### 5.3.2 Biological Relevance

To demonstrate the performance of our method in predicting important domains, it was applied on the FAS-associated death domain protein, FADD\_Human, (PDB Accession number Q13158) which has 208 residues with two domains and one domain-linker located in the interval between 83 and 96 residues according to the Protein Data Bank (RCSB PDB)<sup>1</sup>[116]. Our method succeeded in predicting these two domains as indicated by the orange bars in Figure. 5.7.

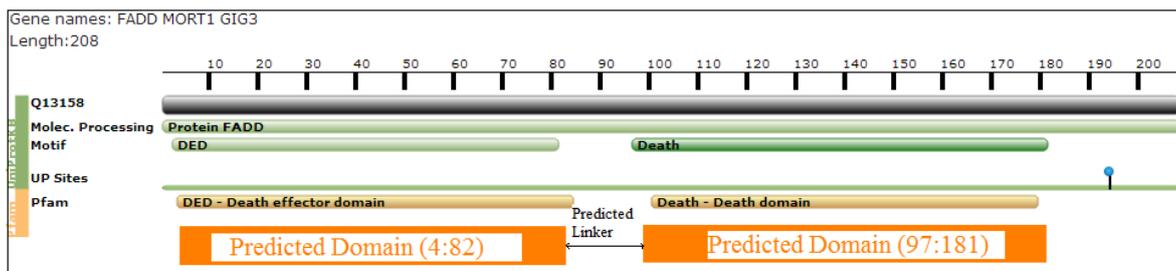


Figure 5.7: FAS-associated death domain protein - Q13158 (FADD\_HUMAN). The protein contains 208 residues and has two domains and a linker according to RCSB-PDB. Our method succeeded in predicting these two domains as indicated by the orange bars.

Another example is illustrated in Figure. 5.8 of the B-lymphocyte antigen CD19 (CD19\_HUMAN). (PDB Accession number P15391) which has 556 residues with two domains and one domain-linker according to the Research Collaboratory for Structural Bioinformatics - Protein Data Bank (RCSB PDB). Our method succeeded in predicting these two immunoglobulin domains as indicated by the orange bars. Immunoglobulin domains may be involved in proteinprotein and proteinligand interactions. The immunoglobulin superfamily domains are involved in the recognition, binding, or adhesion processes of cells. They are commonly associated with roles in the immune system [202].

<sup>1</sup><http://www.rcsb.org/pdb/protein/Q13158>

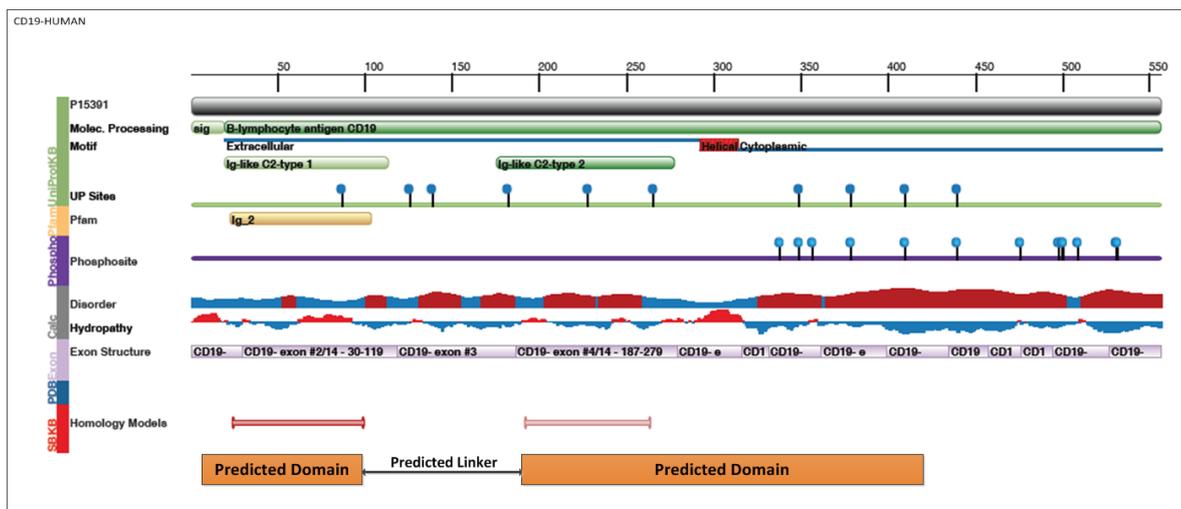


Figure 5.8: B-lymphocyte antigen CD19 - P15391 (CD19\_HUMAN). The protein contains 556 residues and has two domains and a linker according to RCSB-PDB. Our method succeeded in predicting these two domains as indicated by the orange bars.

Figure 5.9 presents the izumo sperm-egg fusion 1, isoform CRA\_c [Homo sapiens] protein which contains 194 residues and has one domain (PF15005) according to NCBI<sup>2</sup>. Our method succeeded in predicting this domain as indicated by the orange bar. The izumo sperm-egg fusion domain is important in fertilization and essential for sperm-egg plasma membrane binding and fusion [203, 204].

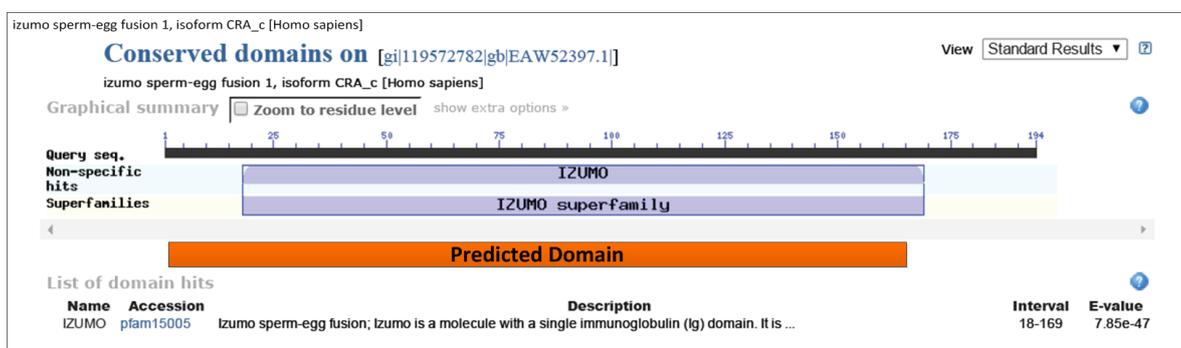


Figure 5.9: Izumo sperm-egg fusion protein. The protein contains 194 residues and has one domain according to NCBI. Our method succeeded in predicting this domain as indicated by the orange bar.

<sup>2</sup><http://www.ncbi.nlm.nih.gov/protein/119572782>

## Chapter 6: PPI Prediction

This chapter presents our third contribution which is predicting protein-protein interactions based on analyzing their interacting structural domains. The method is described in Section 6.1 and experimental results are presented and discussed in Section 6.2.

### 6.1 Method

Following the structural domain identification, we determine that two proteins interact by the means of interacting domain both contain. The validation is done by searching the identified domains in a benchmark domain-domain interaction (DDI) database. This is achieved, as illustrated in Figure 6.1, through the following steps:

- Each of the predicted domains within a given protein pair is searched in the Pfam domain database to find its Pfam ID (Accession Number) by employing the Needleman-Wunsch (NW) global alignment algorithm.
- Based on their Pfam Accession Numbers, domain interactions are searched in three benchmark DDI databases.
- We conclude that two protein interact if they contain one or more interacting domains available from the DDI database.

The details of each step is explained through the proceeding sections.

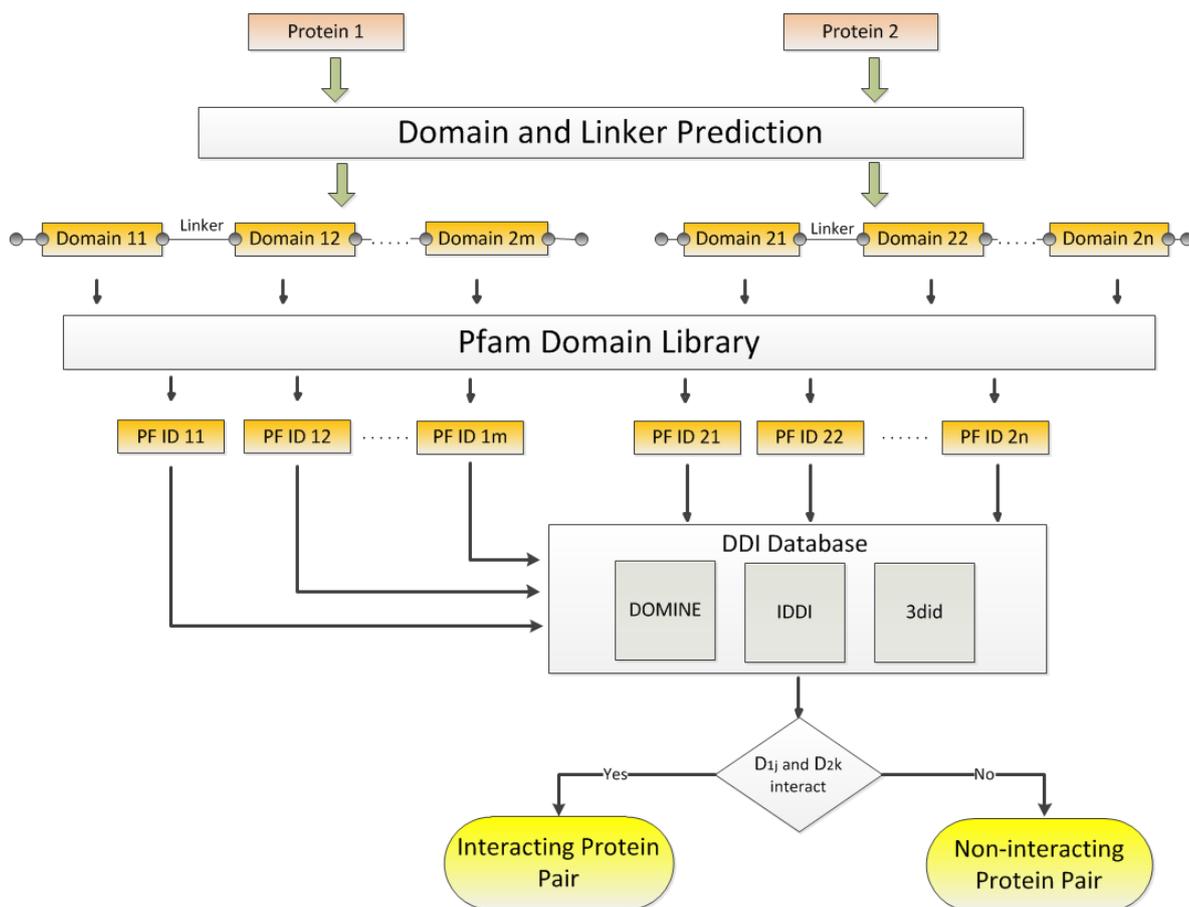


Figure 6.1: Overview of the PPI prediction process.

### 6.1.1 Pfam Search

Each of the predicted domains is searched to find its Pfam Accession Number. This is performed by applying a global sequence alignment of the predicted domain with every entry in Pfam release 27.0 [147] using the Needleman-Wunsch (NW) algorithm [205] and returning the Pfam entry that has the highest alignment score.

Pfam is a large collection of protein families, each represented by multiple sequence alignments and HMMs. The Pfam database consists of two components; Pfam-A and Pfam-B. Pfam-A entries are high quality, manually curated families and cover a large proportion of the sequences in the underlying sequence database. Pfam-B entries are automatically generated and of lower quality and can be useful when no Pfam-A entries are found. We use Pfam-A 27.0 [147] which is the latest Pfam release. Pfam-A

contains 14,830 protein families with 10,626,097 domain entries.

The NW algorithm [205] is a dynamic programming algorithm that measures the similarity score between two sequences by a global gapped alignment and guarantees to find the best alignment. The algorithm provides a method of finding the optimal global alignment of two sequences by maximizing the number of amino acid matches and minimizing the number of gaps necessary to align the two sequences [206].

NW algorithm constructs a two-dimensional matrix in which one of the sequences to be aligned runs down the vertical axis and the other along the horizontal axis. The algorithm finds the best alignment by using optimal alignments of smaller subsequences. The optimal path can then be determined by incremental extension of the optimal subpaths. All possible comparisons between any number of AA pairs are given by pathways through the array and are scored. The alignment is grown from the *C*-terminus towards the *N*-terminus and all possible alignments at each step are rejected except the one with the best score [207]. The NW algorithm consists of three steps; score matrix initialization, matrix filling with maximum scores, and residues traceback for appropriate alignment. NW algorithm is described in Algorithm 2. Regarding its complexity, given two sequences of length  $m$  and  $n$ , the NW algorithm performs the alignment with a time complexity of  $O(mn)$  and a space complexity of  $O(mn)$  [206].

---

**Algorithm 2** Needleman-Wunsch global alignment.

---

**input** two protein sequences  $X$  and  $Y$   
**initialization:**  
 Set  $F(i, 0) = -i.d$  for all  $i = 0, 1, 2, \dots, n$   
 Set  $F(0, j) = -j.d$  for all  $j = 0, 1, 2, \dots, m$   
**for**  $i = 1$  to  $n$  **do: do**  
   **for**  $j = 1$  to  $m$  **do: do**

$$\text{Set } F(i, j) := \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

(6.1)

  Set backtrace  $T(i, j)$  to the maximizing pair  $(i', j')$   
**end for**  
**end for**  
 Score  $\alpha := F(n, m)$   
 Set  $(i, j) := (n, m)$   
**repeat**  
   **if**  $T(i, j) = ((i-1, j-1))$  **then**  
     **print**  $(x_i, y_j)$   
   **else if**  $T(i, j) = ((i-1, j))$  **then**  
     **print**  $(x_i, -)$   
   **else**  
     **print**  $(-, y_j)$   
   **end if**  
   Set  $(i, j) := T(i, j)$   
**until**  $(i, j) = (0, 0)$   
**return** optimal alignment and score  $\alpha$

---

### 6.1.2 DDI Database Search

Domain-Domain Interactions (DDI) occur when two globular domains form a stable interface. The assumption that proteins interact with each other through their domains is widely accepted [89]. Understanding protein interactions at the domain level provides valuable information about binding mechanisms and functional contribution to protein interactions [151]. The initial source of DDI information is the 3D structure of protein complexes but due to the limited availability of 3D structures, DDI prediction methods or their predicted datasets are used as an alternative source [151].

In this work we use three DDI databases; DOMINE, IDDI, and 3did. DOMINE<sup>1</sup> [150, 18] is a database of domain interactions inferred from experimentally characterized high-resolution 3D structures in the Protein Data Bank (PDB)<sup>2</sup>, in addition to predicted domain interactions by thirteen different computational approaches using Pfam domain definitions. DOMINE contains a total of 26,219 DDI pairs among 5,410 domains, out of which 6,634 are inferred from PDB entries, and 21,620 are predicted by at least one computational approach.

The integrated domain-domain interaction analysis system (IDDI)<sup>3</sup> [151] provides 20,4715 unique DDI pairs with different reliability scores. The reliability of the predicted DDI pairs are determined by considering the confidence score of the prediction method, the independence score of the predicted datasets, and the DDI prediction score measured by different prediction methods.

The database of 3D interacting domains (3did)<sup>4</sup> [152] is a collection of 3D struc-

---

<sup>1</sup><http://domine.utdallas.edu>

<sup>2</sup><http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>

<sup>3</sup><http://pcode.kaist.ac.kr/iddi/>

<sup>4</sup><http://3did.irbbarcelona.org>

tures of domain-based interactions in the PDB based on domain definitions from Pfam release 27.0 [147].

## 6.2 Experimental Results and Discussion

To evaluate the performance of our PPI prediction approach, we used a dataset of yeast *Saccharomyces cerevisiae* containing 4,917 protein interaction pairs among 3,713 proteins, and 4,000 randomly-generated non-interacting protein pairs. The data was collected from the DIP [75, 82], Deng *et al.* [131], Schwikowski *et al.* [135]. The dataset of Deng *et al.* is experimentally obtained through two hybrid assays on *Saccharomyces cerevisiae* by Uetz *et al.* [132] and Ito *et al.* [133]. Schwikowski *et al.* gathered their data from yeast two-hybrid, biochemical and genetic data. As non-interacting protein data are unavailable, the negative samples were randomly generated. A protein pair is considered to be a negative sample if the pair does not exist in the interaction set. This dataset was gathered and used by Chen and Liu [89]. Both the positive and negative PPI examples were divided evenly into training and testing datasets. We obtained the domain information from the Pfam-A release 27.0<sup>5</sup> [147].

Once protein domains are identified, our PPI prediction method achieved a prediction accuracy of 97%, sensitivity (recall) of 96%, precision of 98%, and specificity of 98%. The comparisons of our method to the existing PPI prediction approaches are summarized in Figure 6.2 which clearly shows that the proposed method outperformed the existing PPI predictors in terms of sensitivity and specificity.

In terms of the prediction performance of the whole process of domain identification and PPI prediction, we achieved a prediction accuracy of 78%, sensitivity of 60%, precision of 94%, and specificity of 96%. This reduction in prediction performance is due to the fact that some of the predicted domains in few proteins are either shorter or longer than the actual domains or the fact that our method sometimes predicts several

---

<sup>5</sup><http://pfam.sanger.ac.uk>

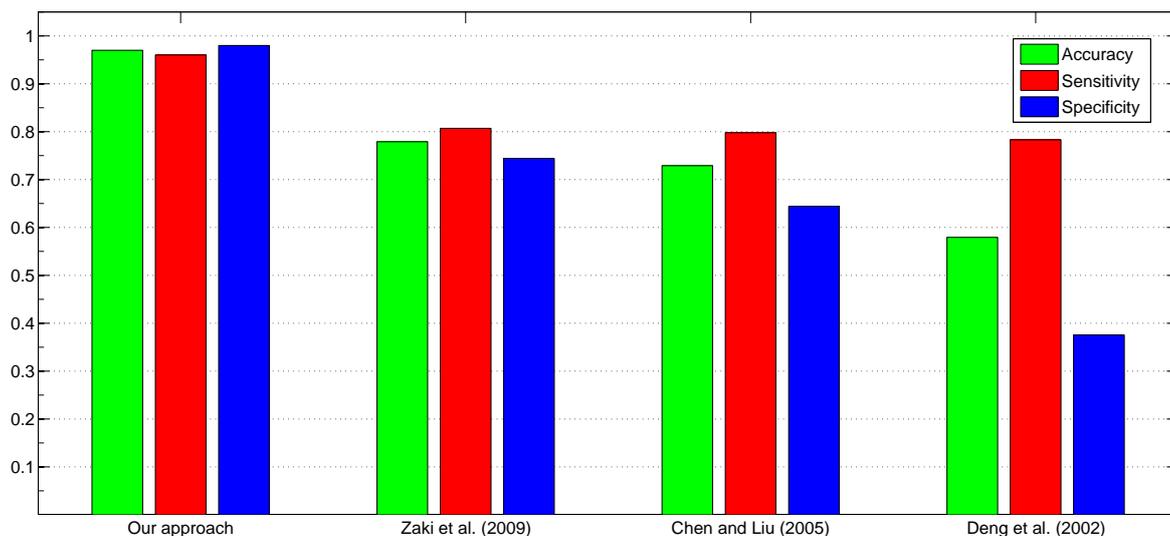


Figure 6.2: Accuracy, sensitivity, and specificity of the state-of-the-art PPI predictors compared to our approach.

short domains in a location that contains a long actual domain, and therefore, these predicted domains do not exactly match with domains in the Pfam database. To overcome this issue, we followed DomCut [4] where they consider the domain linker to be in the range of 10 and 100, and thus we extended the domain prediction stage by adding a post-processing step where if several adjacent domains are identified and they are a part by less than 10 AA residues, they will be concatenated into a single domain. As a result, the overall prediction accuracy is improved to 90%.

Although this approach achieved very high PPI prediction accuracy, the PPI prediction performance is strongly dependent on domain prediction accuracy and if domains are not accurately identified, PPI prediction will be negatively affected. One of the limitations of this approach is the computational time of the sequence alignment step as the NW algorithm is applied to calculate the alignment score for each identified domain against all the 10,626,097 Pfam domain entries.

To demonstrate the effectiveness of the proposed method in identifying domains and predicting protein interactions, let us take YCR077C and YDL160C as an example of interacting protein pairs according to our benchmark dataset. As shown in Figure 6.3,

two domains are identified in the first protein in the regions [1-224] and [241-788] and two domains are identified in the second protein in the regions [71-235] and [303-378]. The Pfam accession number for these domains are PF09770, PF09770, PF00270, and PF00271, respectively. When these domains are searched through the DDI databases, it is confirmed by 3did that PF09770 interacts with PF00270. As a result, the model reports that the two proteins interact.

Similarly, YDR477W and YER027C represent another example of interacting protein pairs. As shown in Figure 6.4, two domains are identified in the first protein in the regions [55-306] and [344-389] and one domain is identified in the second protein in the region [306-415]. The Pfam accession number for the two domains of the first protein are PF00069 and PF08587, and the Pfam accession number of the domain in the second protein is PF04739. When these domains are searched through the DDI databases, it is confirmed by IDDI that both PF00069 and PF08587 interact with PF04739 and retrieved by 3did that PF08587 interacts with PF04739. As a result, the model reports that the two proteins interact.

YDR044W and YCR014C represent an example of non-interacting protein pairs. As shown in Figure 6.5, one domain is identified in the first protein in the region [14-327] and three domains are identified in the second protein in the regions [188-253], [326-407], and [517-574]. The Pfam accession number for the domain of the first protein is PF01218 and the Pfam accession number for three domains in the second protein are PF14716, PF14792, and PF14791. When these domains are searched through the DDI databases, no interacting domains were found. As a result, the model reports that the two proteins are not interacting.

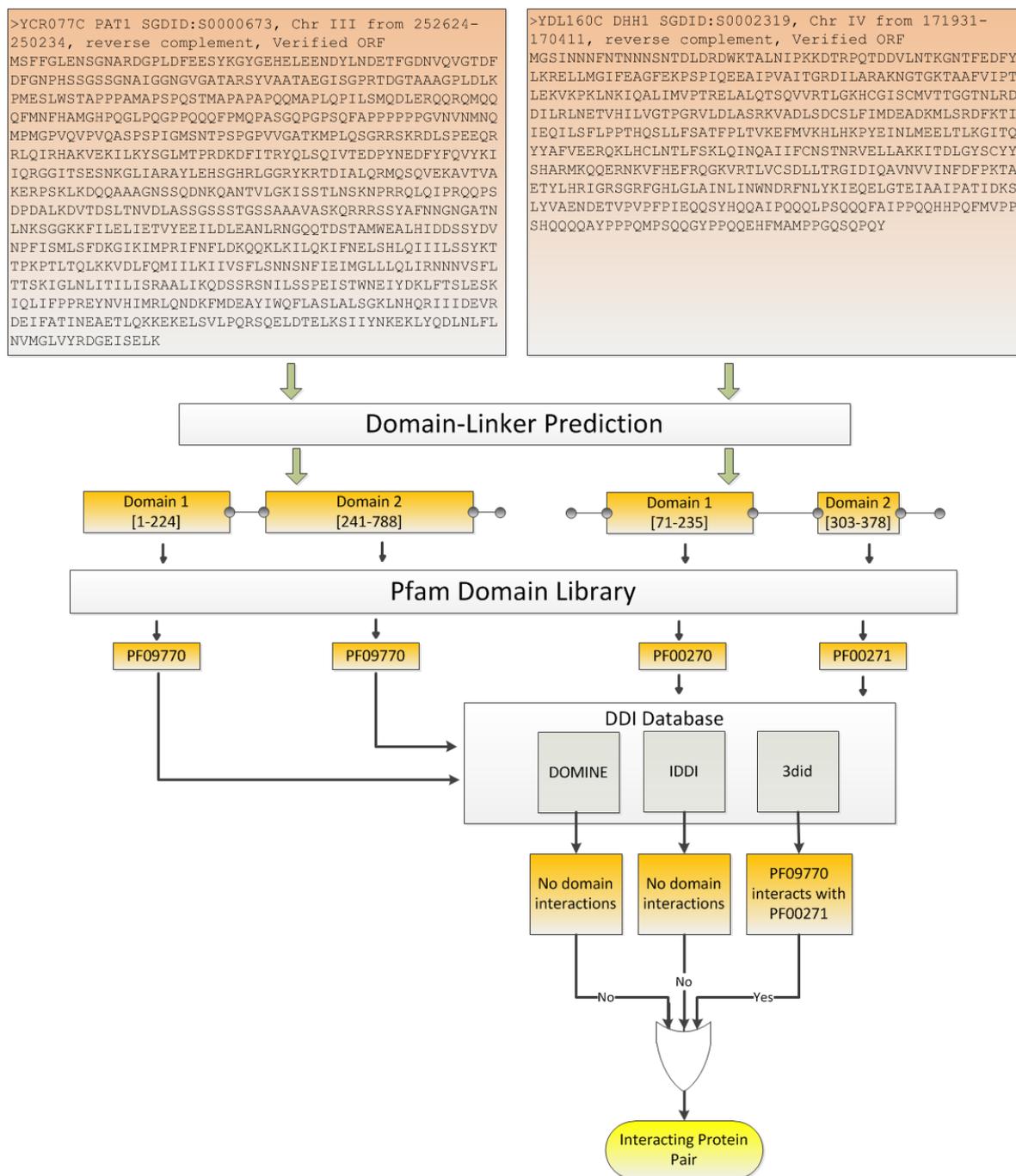


Figure 6.3: PPI prediction for YCR077C and YDL160C proteins. Two domains are identified in the first protein and two domains are identified in the second protein. 3did retrieves that domain PF09770 from the first protein interacts with domain PF00270 from the second protein. As a result, the model predicts that the two proteins are interacting.

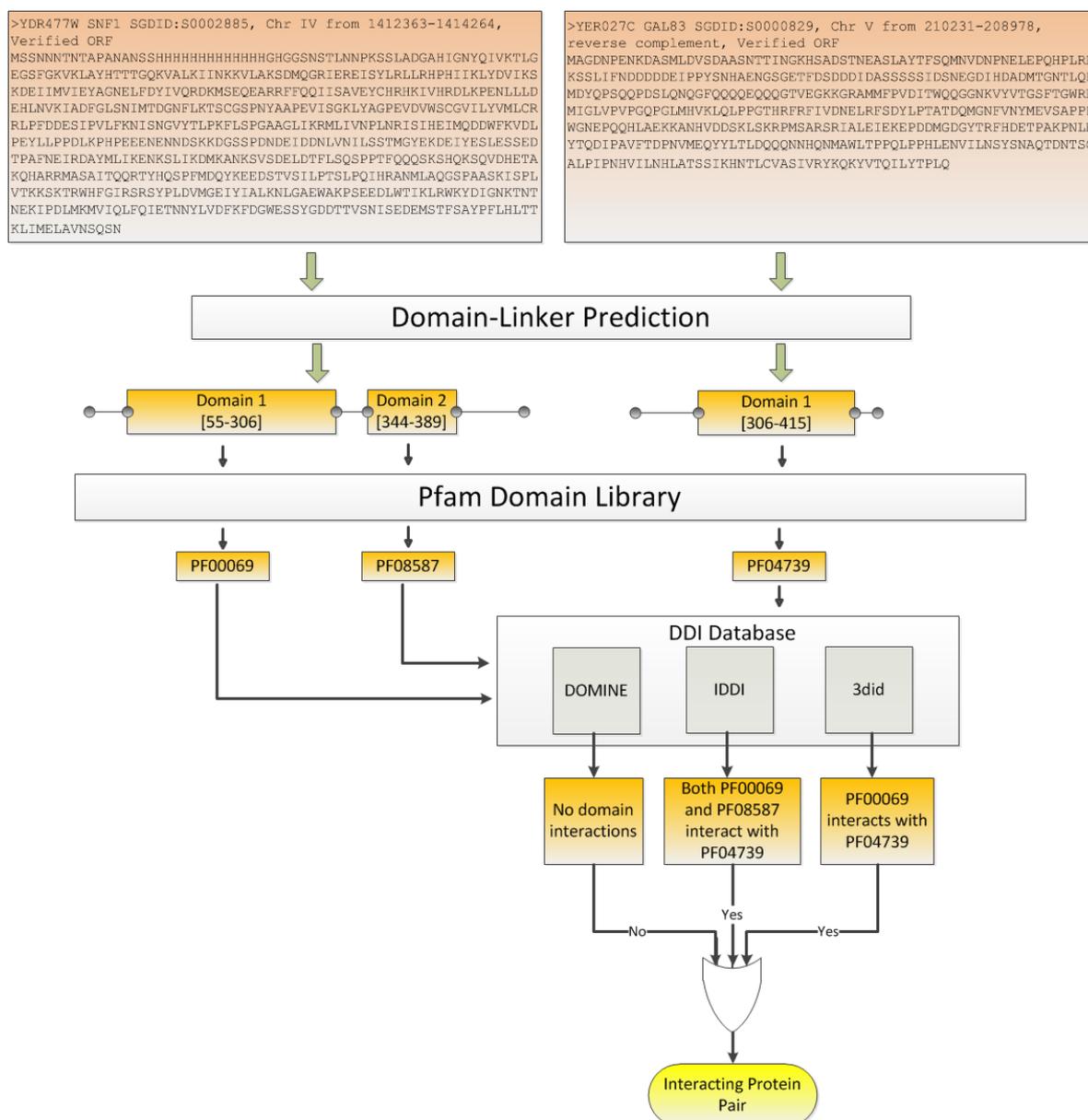


Figure 6.4: PPI prediction for YDR477W and YER027C proteins. Two domains are identified in the first protein and one domain is identified in the second protein. Domain interactions are retrieved in 3did and IDDI. As a result, the model predicts that the two proteins are interacting.

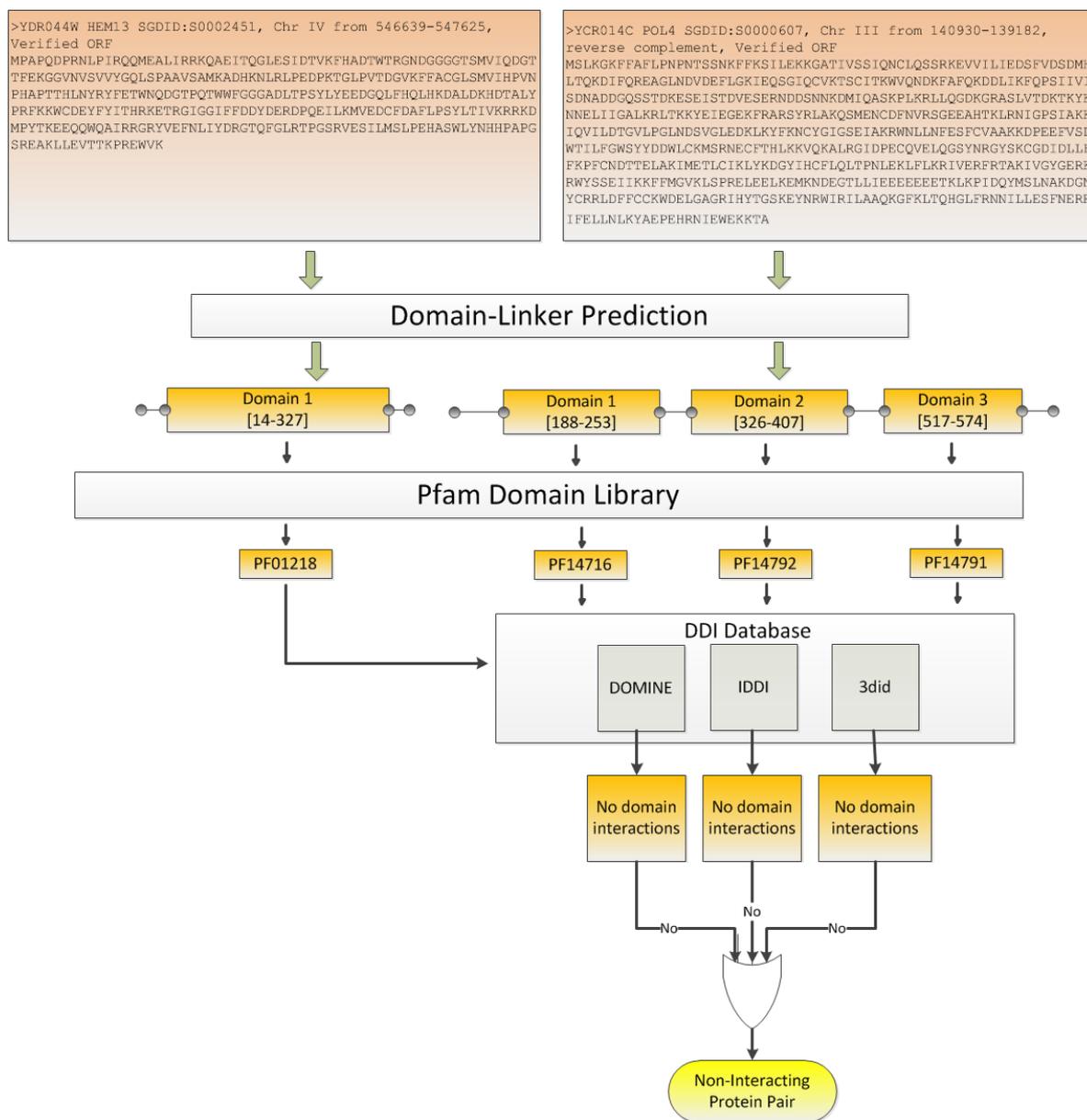


Figure 6.5: PPI prediction for YDR044W and YCR014C proteins. One domain is identified in the first protein and three domains are identified in the second protein. No domain interactions are retrieved in the DDI database. As a result, the model predicts that the two proteins are not interacting.

## Chapter 7: Conclusions and Future Work

### Directions

In this research work, we employ structural domain and inter-domain linker prediction into predicting PPIs. We propose a novel method for predicting inter-domain linkers within proteins. This is achieved through introducing the concept of AA compositional index. The linker knowledge, represented by AA compositional index, is then enhanced by biological knowledge through combining it with AA physiochemical properties to develop a Random Forest classifier for predicting novel domains and linkers. Following the structural domain identification step, we predict whether two proteins interact or not by analyzing the interacting structural domains that they contain.

The three main contributions of this work can be summarized as follows; In the first contribution, we developed CISA as a method for detecting protein domain-linker regions based on AA compositional index and Simulated Annealing. Experimental results showed that this method outperformed the currently available approaches of domain-linker prediction in terms of recall, precision, and F1-score. It was also shown that CISA is capable of predicting novel linkers which could lead to the identification of crucial structural domains such as RING-finger and carboxy-terminal domains. The main reasons behind the considerable accuracy achieved by CISA is the improvement in the concept of AA compositional index and the adoption of the SA algorithm to refine the prediction by finding the optimal set of threshold values that separate domains from

linker regions. CISA has a potential to perform well if it is applied to human proteins where novel domain linkers could be recognized.

Although SA has significantly improved the prediction, additional tuning could accomplish more effective and flexible prediction. One of these tuning strategies is the use of dynamic chunk sizes which could, in turn, obtain better optimization and more accurate prediction. This work can be extended by exploring other compositional index models such as the weighted sum or the weighted product of linker index and AA composition, and employ SA to find the optimal weights along with the optimal threshold that separate linker regions from domain regions. Furthermore, other optimization techniques such as Genetic Algorithm can be examined and compared to SA in domain linker prediction, or both techniques could be combined in a hybrid approach.

In the second contribution, we developed a novel machine-learning approach to predict novel domains and linkers. This is achieved by combining the compositional index with AA physiochemical properties to construct a novel protein profile. A sliding window technique is applied to extract and normalize the AA features and takes into consideration the dependences of each AA with its neighborhood. Then, a well-optimized Random Forest domain-linker classifier is constructed and trained by these protein features. The utility of the proposed approach is illustrated on two well-known benchmark datasets by achieving a high prediction accuracy and outperforming the state-of-the-art domain predictors in terms of recall, precision, and F1-score. The proposed approach successfully eliminates some of the data pre-processing steps such as class weights or data re-sampling techniques, and proves that the model can handle imbalanced data and is not biased towards the majority class.

Although various ML-based domain prediction approaches have been developed, they have shown a limited capability in multi-domain protein prediction. Capturing

long-term AA dependencies and developing a more suitable representation of protein sequence profiles that includes evolutionary information may lead to better model performance. Existing approaches showed a limited capability in exploiting long-range interactions that exist among amino acids and participate in the formation of protein secondary structures. Residues can be adjacent in 3D space while located far apart in the AA sequence. [9, 33].

Regarding protein sequence profile representation, the proposed input profiles in most domain-linker predictors still provides insufficient structural information to reach the maximum accuracy. One reason behind the limited capability of multi-domain protein predictors is the disagreement of domain assignment within different protein databases. The agreement between domain databases covers about 80% of single domain proteins and about 66% of multi-domain proteins only [34]. This disagreement is due to the variance in the experimental methods used in domain assignment. The most predominant techniques used to experimentally determine protein 3D structures are X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). However, their conformational results of domain assignment vary in about 20% so that the upper limit accuracy for such domain-linker prediction task could be about 80%.

This approach can be extended by examining longer averaging window sizes in order to capture long-range interactions that exist among amino acids and participate in the formation of protein secondary and tertiary structures. Residues can be adjacent in 3D space while located far apart in the AA sequence. The averaging window formula can also be improved to a weighted average so that the closer AA neighbors to the central residue can take higher weights than farther ones. Although the proposed approach successfully handles the imbalanced protein data, data balancing techniques such as re-sampling can be integrated and tested for further improvement of the model per-

formance. Comparing the performance of RF in domain prediction with other ensemble methods such as bagging and boosting is one of the future work directions. Emerging ensemble methods such as ensemble of support vector machines, meta-ensemble, and ensemble of heterogeneous classification algorithms are promising directions.

In the third contribution, we developed a novel PPI prediction approach based on characterizing structural domains within proteins and analyzing their interactions. Each of the predicted domains within a given protein pair is searched in the Pfam domain database to find its Pfam Accession Number by employing the Needleman-Wunsch (NW) global alignment algorithm. Based on their Pfam Accession Numbers, domain interactions are searched in three benchmark domain-domain interaction databases. We determine that two proteins interact if a domain in the first protein is interacting with a domain in the second protein as confirmed by at least one of the benchmark DDI databases. When tested on a dataset of *Saccharomyces cerevisiae* protein pairs, the method showed a very high capability of predicting PPIs outperforming several existing predictors. One of our future goals is to develop a web server that enables users to enter a protein pairs and return their structural domains and whether they are interacting or not.

One of the limitations of this approach is the computational time of the sequence alignment step as the NW algorithm is applied to calculate the alignment score for each identified domain against each of the Pfam domain entries. Therefore, the NW alignment can be a further research area for parallel computing. Although this approach achieved very high PPI prediction accuracy, the PPI prediction performance is strongly dependent on domain prediction performance. If domains are not accurately identified, PPI prediction will be negatively affected. Therefore, any improvement in our previous contributions of domain and linker prediction can lead to improvement in PPI predic-

tion. One of the possible future directions is to include more DDI databases in order to have better validation and to search and include validated non-DDI databases to validate non-interacting protein pairs.

## Bibliography

- [1] Taniguchi, N., “Amino acids and proteins,” *Medical Biochemistry, Ed*, vol. 3, pp. 5–21, 2010.
- [2] Petsko, G. A. and Ringe, D., *Protein structure and function*. New Science Press, 2004.
- [3] Zhu, H., Domingues, F. S., Sommer, I., and Lengauer, T., “Noxclass: prediction of protein-protein interaction types,” *BMC bioinformatics*, vol. 7, no. 1, p. 27, 2006.
- [4] Suyama, M. and Ohara, O., “Domcut: prediction of inter-domain linker regions in amino acid sequences,” *Bioinformatics*, vol. 19, no. 5, pp. 673–674, 2003.
- [5] Zaki, N., Bouktif, S., and Lazarova-Molnar, S., “A genetic algorithm to enhance transmembrane helices prediction,” in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pp. 347–354, ACM, 2011.
- [6] Tanaka, T., Yokoyama, S., and Kuroda, Y., “Improvement of domain linker prediction by incorporating loop-length-dependent characteristics,” *Peptide Science*, vol. 84, no. 2, pp. 161–168, 2006.
- [7] Ebina, T., Toh, H., and Kuroda, Y., “Loop-length-dependent svm prediction of domain linkers for high-throughput structural proteomics,” *Peptide Science*, vol. 92, no. 1, pp. 1–8, 2009.
- [8] Chothia, C., “Proteins. one thousand families for the molecular biologist.,” *Nature*, vol. 357, no. 6379, p. 543, 1992.
- [9] Yoo, P. D., Sikder, A. R., Taheri, J., Zhou, B. B., and Zomaya, A. Y., “Domnet: protein domain boundary prediction using enhanced general regression network and new profiles,” *NanoBioscience, IEEE Transactions on*, vol. 7, no. 2, pp. 172–181, 2008.
- [10] Hondoh, T., Kato, A., Yokoyama, S., and Kuroda, Y., “Computer-aided nmr assay for detecting natively folded structural domains,” *Protein science*, vol. 15, no. 4, pp. 871–883, 2006.
- [11] Bhaskara, R. M., de Brevern, A. G., and Srinivasan, N., “Understanding the role of domain–domain linkers in the spatial orientation of domains in multi-domain proteins,” *Journal of Biomolecular Structure and Dynamics*, no. ahead-of-print, pp. 1–14, 2012.
- [12] Xenarios, I. and Eisenberg, D., “Protein interaction databases,” *Current Opinion in Biotechnology*, vol. 12, no. 4, pp. 334–339, 2001.

- [13] Kim, W. K., Park, J., Suh, J. K., *et al.*, “Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair,” *Genome Informatics Series*, pp. 42–50, 2002.
- [14] Zaki, N., Lazarova-Molnar, S., El-Hajj, W., and Campbell, P., “Protein-protein interaction based on pairwise similarity,” *BMC bioinformatics*, vol. 10, no. 1, p. 150, 2009.
- [15] Apic, G., Gough, J., and Teichmann, S. A., “Domain combinations in archaeal, eubacterial and eukaryotic proteomes,” *Journal of molecular biology*, vol. 310, no. 2, pp. 311–325, 2001.
- [16] Pawson, T. and Nash, P., “Assembly of cell regulatory systems through protein interaction domains,” *science*, vol. 300, no. 5618, pp. 445–452, 2003.
- [17] Shoemaker, B., Panchenko, A., and Bryant, S., “Finding biologically relevant protein domain interactions: Conserved binding mode analysis,” *PROTEIN SCIENCE*, vol. 15, pp. 352–361, FEB 2006.
- [18] Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R., “Domine: a comprehensive collection of known and predicted domain-domain interactions,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D730–D735, 2011.
- [19] Berg, J. M., Tymoczko, J. L., and Stryer, L., “Biochemistry,” 2002.
- [20] Berg, J. M., Tymoczko, J. L., and Stryer, L., “Protein structure and function,” 2002.
- [21] Bader, G. D. and Hogue, C. W., “Analyzing yeast protein-protein interaction data obtained from different sources,” *Nat Biotech*, vol. 20, pp. 991–997, OCT 2002.
- [22] Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P., “Comparative assessment of large-scale data sets of protein-protein interactions,” *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.
- [23] Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J., “Evaluation of different biological data and computational classification methods for use in protein interaction prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.
- [24] Chou, K.-C., “Some remarks on protein attribute prediction and pseudo amino acid composition,” *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011.
- [25] Liu, J. and Rost, B., “Sequence-based prediction of protein domains,” *Nucleic acids research*, vol. 32, no. 12, pp. 3522–3530, 2004.
- [26] Cheng, J., Sweredoski, M. J., and Baldi, P., “Dompro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks,” *Data Mining and Knowledge Discovery*, vol. 13, no. 1, pp. 1–10, 2006.
- [27] Gewehr, J. E. and Zimmer, R., “Ssep-domain: protein domain prediction by alignment of secondary structure elements and profiles,” *Bioinformatics*, vol. 22, no. 2, pp. 181–187, 2006.
- [28] Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., *et al.*, “Structure-based prediction

- of protein-protein interactions on a genome-wide scale,” *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.
- [29] Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P., “Scratch: a protein structure and structural feature prediction server,” *Nucleic acids research*, vol. 33, no. suppl 2, pp. W72–W76, 2005.
- [30] Sim, J., Kim, S.-Y., and Lee, J., “Pprodo: prediction of protein domain boundaries using neural networks,” *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 3, pp. 627–632, 2005.
- [31] Ebina, T., Toh, H., and Kuroda, Y., “Drop: an svm domain linker predictor trained with optimal features selected by random forest,” *Bioinformatics*, vol. 27, no. 4, pp. 487–494, 2011.
- [32] Melo, J. C., Cavalcanti, G., and Guimaraes, K., “Pca feature extraction for protein structure prediction,” in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4, pp. 2952–2957, IEEE, 2003.
- [33] Chen, J. and Chaudhari, N. S., “Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction,” *Soft Computing*, vol. 10, no. 4, pp. 315–324, 2006.
- [34] Marsden, R. L., McGuffin, L. J., and Jones, D. T., “Rapid protein domain assignment from amino acid sequence using predicted secondary structure,” *Protein Science*, vol. 11, no. 12, pp. 2814–2824, 2002.
- [35] Pietzsch, J., “Protein folding technology,” *Protein Folding and Disease*, 2007.
- [36] Wüthrich, K., “Nuclear magnetic resonance (nmr) spectroscopy of proteins,” *eLS*, 2001.
- [37] Krishnan, V. and Rupp, B., “Macromolecular structure determination: Comparison of x-ray crystallography and nmr spectroscopy,” *eLS*, 2001.
- [38] Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J., “Globplot: exploring protein sequences for globularity and disorder,” *Nucleic acids research*, vol. 31, no. 13, pp. 3701–3708, 2003.
- [39] Dong, Q., Wang, X., Lin, L., and Xu, Z., “Domain boundary prediction based on profile domain linker propensity index,” *Computational biology and chemistry*, vol. 30, no. 2, pp. 127–133, 2006.
- [40] Pang, C. N., Lin, K., Wouters, M. A., Heringa, J., and George, R. A., “Identifying foldable regions in protein sequence from the hydrophobic signal,” *Nucleic acids research*, vol. 36, no. 2, pp. 578–588, 2008.
- [41] Udworthy, D. W., Merski, M., and Townsend, C. A., “A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type i polyketide synthase,” *Journal of molecular biology*, vol. 323, no. 3, pp. 585–598, 2002.
- [42] Dumontier, M., Yao, R., Feldman, H. J., and Hogue, C. W., “Armadillo: domain boundary prediction by amino acid composition,” *Journal of molecular biology*, vol. 350, no. 5, pp. 1061–1073, 2005.
- [43] Walsh, I., Martin, A. J., Mooney, C., Rubagotti, E., Vullo, A., and Pollastri, G., “Ab initio and homology based prediction of protein domains by recursive neural networks,” *BMC bioinformatics*, vol. 10, no. 1, p. 195, 2009.

- [44] Eickholt, J., Deng, X., and Cheng, J., “Dobo: Protein domain boundary prediction by integrating evolutionary signals and machine learning,” *BMC bioinformatics*, vol. 12, no. 1, p. 43, 2011.
- [45] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C., “Scop: a structural classification of proteins database for the investigation of sequences and structures,” *Journal of molecular biology*, vol. 247, no. 4, pp. 536–540, 1995.
- [46] Specht, D. F., “A general regression neural network,” *Neural Networks, IEEE Transactions on*, vol. 2, no. 6, pp. 568–576, 1991.
- [47] George, R. A., Lin, K., and Heringa, J., “Scooby-domain: prediction of globular domains in protein sequence,” *Nucleic acids research*, vol. 33, no. suppl 2, pp. W160–W163, 2005.
- [48] Bondugula, R., Lee, M. S., and Wallqvist, A., “Fiefdom: a transparent domain boundary recognition system using a fuzzy mean operator,” *Nucleic acids research*, vol. 37, no. 2, pp. 452–462, 2009.
- [49] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al.*, “The pfam protein families database,” *Nucleic acids research*, p. gkr1065, 2011.
- [50] Xue, Z., Xu, D., Wang, Y., and Zhang, Y., “Threadom: extracting protein domain boundary information from multiple threading alignments,” *Bioinformatics*, vol. 29, no. 13, pp. i247–i256, 2013.
- [51] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., “Gapped blast and psi-blast: a new generation of protein database search programs,” *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [52] Fischer, D., Barret, C., Bryson, K., Eloffson, A., Godzik, A., Jones, D., Karplus, K. J., Kelley, L. A., MacCallum, R. M., Pawowski, K., *et al.*, “Cafasp-1: critical assessment of fully automated structure prediction methods,” *Proteins: Structure, Function, and Bioinformatics*, vol. 37, no. S3, pp. 209–217, 1999.
- [53] Saini, H. K. and Fischer, D., “Meta-dp: domain prediction meta-server,” *Bioinformatics*, vol. 21, no. 12, pp. 2917–2920, 2005.
- [54] Bairoch, A. and Apweiler, R., “The swiss-prot protein sequence database and its supplement trembl in 2000,” *Nucleic acids research*, vol. 28, no. 1, pp. 45–48, 2000.
- [55] Simossis, V. and Heringa, J., “Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information,” *Nucleic Acids Research*, vol. 33, no. suppl 2, pp. W289–W294, 2005.
- [56] Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., and Thornton, J. M., “Cath—a hierarchic classification of protein domain structures,” *Structure*, vol. 5, no. 8, pp. 1093–1109, 1997.
- [57] Pearl, J., “Heuristics: intelligent search strategies for computer problem solving,” 1984.
- [58] Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd ed., 2009.

- [59] Keller, J. M., Gray, M. R., and Givens, J. A., “A fuzzy k-nearest neighbor algorithm,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 580–585, 1985.
- [60] Wu, S. and Zhang, Y., “Lomets: a local meta-threading-server for protein structure prediction,” *Nucleic acids research*, vol. 35, no. 10, pp. 3375–3382, 2007.
- [61] Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., and Basu, D. K., “Improved prediction of multi-domains in protein chains using a support vector machine,” *International Journal of Recent Trends in Engineering*, vol. 2, no. 3, 2009.
- [62] Li, B.-Q., Hu, L.-L., Chen, L., Feng, K.-Y., Cai, Y.-D., and Chou, K.-C., “Prediction of protein domain with mrmr feature selection and analysis,” *PLoS ONE*, vol. 7, no. 6, p. e39308, 2012.
- [63] Consortium, U. *et al.*, “The universal protein resource (uniprot) in 2010,” *Nucleic acids research*, vol. 38, no. suppl 1, pp. D142–D148, 2010.
- [64] Bartel, P. L. and Fields, S., *The yeast two-hybrid system*. Oxford University Press, 1997.
- [65] Gavin, A.-C., Bösch, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., *et al.*, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.
- [66] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B., “A generic protein purification method for protein complex characterization and proteome exploration,” *Nature biotechnology*, vol. 17, no. 10, pp. 1030–1032, 1999.
- [67] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., *et al.*, “Global analysis of protein activities using proteome chips,” *science*, vol. 293, no. 5537, pp. 2101–2105, 2001.
- [68] Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., *et al.*, “A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules,” *Science*, vol. 295, no. 5553, pp. 321–324, 2002.
- [69] Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H., “Predicting protein–protein interactions based only on sequences information,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.
- [70] Szilágyi, A., Grimm, V., Arakaki, A. K., and Skolnick, J., “Prediction of physical protein–protein interactions,” *Physical biology*, vol. 2, no. 2, p. S1, 2005.
- [71] Porollo, A. and Meller, J., “Computational methods for prediction of protein–protein interaction sites,” *Protein-Protein Interactions-Computational and Experimental Tools; W. Cai and H. Hong, Eds. InTech*, vol. 472, pp. 3–26, 2012.

- [72] Pazos, F. and Valencia, A., “Similarity of phylogenetic trees as indicator of protein–protein interaction,” *Protein engineering*, vol. 14, no. 9, pp. 609–614, 2001.
- [73] McLachlan, A. D., “Tests for comparing related amino-acid sequences. cytochrome *c* and cytochrome *c*<sub>551</sub>,” *Journal of molecular biology*, vol. 61, no. 2, pp. 409–424, 1971.
- [74] Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N., *et al.*, “Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs,” *BMC bioinformatics*, vol. 7, no. 1, p. 365, 2006.
- [75] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D., “The database of interacting proteins: 2004 update,” *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.
- [76] Mewes, H.-W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B., “Mips: a database for genomes and protein sequences,” *Nucleic acids research*, vol. 30, no. 1, pp. 31–34, 2002.
- [77] Pitre, S., North, C., Alamgir, M., Jessulat, M., Chan, A., Luo, X., Green, J., Dumontier, M., Dehne, F., and Golshani, A., “Global investigation of protein–protein interactions in yeast *saccharomyces cerevisiae* using re-occurring short polypeptide sequences,” *Nucleic acids research*, vol. 36, no. 13, pp. 4286–4294, 2008.
- [78] Liu, C. H., Li, K.-C., and Yuan, S., “Human protein–protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence,” *Bioinformatics*, vol. 29, no. 1, pp. 92–98, 2013.
- [79] Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K. O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F., *et al.*, “Evola: Ortholog database of all human genes in h-invdB with manual curation of phylogenetic trees,” *Nucleic acids research*, vol. 36, no. suppl 1, pp. D787–D792, 2008.
- [80] Prasad, T. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.*, “Human protein reference database-2009 update,” *Nucleic acids research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.
- [81] Guo, Y., Yu, L., Wen, Z., and Li, M., “Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences,” *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.
- [82] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D., “Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions,” *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.
- [83] Cristianini, N. and Shawe-Taylor, J., *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

- [84] Zaki, N., Wolfsheimer, S., Nuel, G., and Khuri, S., “Conotoxin protein classification using free scores of words and support vector machines,” *BMC bioinformatics*, vol. 12, no. 1, p. 217, 2011.
- [85] Vapnik, V. N., “Statistical learning theory. adaptive and learning systems for signal processing, communications, and control,” *Simon Haykin*, 1998.
- [86] Smith, T. F. and Waterman, M. S., “Identification of common molecular subsequences,” *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [87] Zaki, N., Deris, S., and Alashwal, H., “Protein-protein interaction detection based on substring sensitivity measure,” *International journal of biomedical sciences*, vol. 2, no. 1, pp. 148–154, 2006.
- [88] Zaki, N., “Protein-protein interaction prediction using homology and inter-domain linker region information,” *Advances in Electrical Engineering and Computational Science, Lecture Notes in Electrical Engineering*, vol. 39, pp. 635–645, 2009.
- [89] Chen, X.-W. and Liu, M., “Prediction of protein–protein interactions using random decision forest framework,” *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.
- [90] Liao, L. and Noble, W. S., “Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships,” *Journal of computational biology*, vol. 10, no. 6, pp. 857–868, 2003.
- [91] Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T., “Protein homology detection using string alignment kernels,” *Bioinformatics*, vol. 20, no. 11, pp. 1682–1689, 2004.
- [92] Roy, S., Martinez, D., Platero, H., Lane, T., and Werner-Washburne, M., “Exploiting amino acid composition for predicting protein-protein interactions,” *PloS ONE*, vol. 4, no. 11, p. e7813, 2009.
- [93] Gomez, S. M., Noble, W. S., and Rzhetsky, A., “Learning to predict protein–protein interactions from protein sequences,” *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.
- [94] Martin, S., Roe, D., and Faulon, J.-L., “Predicting protein–protein interactions using signature products,” *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.
- [95] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M., “Biogrid: a general repository for interaction datasets,” *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.
- [96] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., *et al.*, “A map of the interactome network of the metazoan *c. elegans*,” *Science*, vol. 303, no. 5657, pp. 540–543, 2004.
- [97] Yu, C.-Y., Chou, L.-C., and Chang, D. T., “Predicting protein-protein interactions in unbalanced data using the primary structure of proteins,” *BMC bioinformatics*, vol. 11, no. 1, p. 167, 2010.
- [98] Oyang, Y.-J., Hwang, S.-C., Ou, Y.-Y., Chen, C.-Y., and Chen, Z.-W., “Data classification with radial basis function networks based on a novel kernel

- density estimation algorithm,” *Neural Networks, IEEE Transactions on*, vol. 16, no. 1, pp. 225–236, 2005.
- [99] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T., Gronborg, M., *et al.*, “Development of human protein reference database as an initial platform for approaching systems biology in humans,” *Genome research*, vol. 13, no. 10, pp. 2363–2371, 2003.
- [100] Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., *et al.*, “Human protein reference database2006 update,” *Nucleic acids research*, vol. 34, no. suppl 1, pp. D411–D414, 2006.
- [101] Valente, G. T., Acencio, M. L., Martins, C., and Lemke, N., “The development of a universal in silico predictor of protein-protein interactions,” *PLoS ONE*, vol. 8, no. 5, p. e65587, 2013.
- [102] Kern, C., Gonzalez, A. J., Liao, L., and Vijay-Shanker, K., “Predicting interacting residues using long-distance information and novel decoding in hidden markov models,” *Nanoscience, IEEE Transactions on*, vol. 12, no. 13, pp. 158–164, 2013.
- [103] Friedrich, T., Pils, B., Dandekar, T., Schultz, J., and Müller, T., “Modelling interaction sites in protein domains with interaction profile hidden markov models,” *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.
- [104] Stein, A., Russell, R. B., and Aloy, P., “3did: interacting protein domains of known three-dimensional structure,” *Nucleic acids research*, vol. 33, no. suppl 1, pp. D413–D417, 2005.
- [105] Rabiner, L. and Juang, B.-H., “An introduction to hidden markov models,” *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.
- [106] Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D., “Hidden markov models in computational biology: Applications to protein modeling,” *Journal of molecular biology*, vol. 235, no. 5, pp. 1501–1531, 1994.
- [107] Eddy, S. R., “Profile hidden markov models,” *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.
- [108] Krogh, A. *et al.*, “An introduction to hidden markov models for biological sequences,” *New Comprehensive Biochemistry*, vol. 32, pp. 45–63, 1998.
- [109] Yoon, B.-J., “Hidden markov models and their applications in biological sequence analysis,” *Current genomics*, vol. 10, no. 6, p. 402, 2009.
- [110] Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O., “Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism,” *Nature protocols*, vol. 6, no. 9, pp. 1341–1354, 2011.
- [111] Hubbard, S. J. and Thornton, J. M., “Naccess,” *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, vol. 2, no. 1, 1993.
- [112] Shatsky, M., Nussinov, R., and Wolfson, H. J., “A method for simultaneous alignment of multiple protein structures,” *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 1, pp. 143–156, 2004.

- [113] Mashiach, E., Nussinov, R., and Wolfson, H. J., “Fiberdock: flexible induced-fit backbone refinement in molecular docking,” *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 6, pp. 1503–1519, 2010.
- [114] Ozbabacan, S. E. A., Keskin, O., Nussinov, R., and Gursoy, A., “Enriching the human apoptosis pathway by predicting the structures of protein–protein complexes,” *Journal of structural biology*, vol. 179, no. 3, pp. 338–346, 2012.
- [115] Tuncbag, N., Kar, G., Gursoy, A., Keskin, O., and Nussinov, R., “Towards inferring time dimensionality in protein–protein interaction networks by integrating structures: the p53 example,” *Molecular BioSystems*, vol. 5, no. 12, pp. 1770–1778, 2009.
- [116] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E., “The protein data bank,” *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.
- [117] Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B. M., Eramian, D., *et al.*, “Modbase: a database of annotated comparative protein structure models and associated resources,” *Nucleic acids research*, vol. 34, no. suppl 1, pp. D291–D295, 2006.
- [118] Mirkovic, N., Li, Z., Parnassa, A., and Murray, D., “Strategies for high-throughput comparative modeling: Applications to leverage analysis in structural genomics and protein family organization,” *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 766–777, 2007.
- [119] Henrick, K. and Thornton, J. M., “Pqs: a protein quaternary structure file server,” *Trends in biochemical sciences*, vol. 23, no. 9, pp. 358–361, 1998.
- [120] Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., *et al.*, “The interpro database, an integrated documentation resource for protein families, domains and functional sites,” *Nucleic acids research*, vol. 29, no. 1, pp. 37–40, 2001.
- [121] Han, D., Kim, H.-S., Seo, J., and Jang, W., “A domain combination based probabilistic framework for protein–protein interaction prediction,” *GENOME INFORMATICS SERIES*, pp. 250–260, 2003.
- [122] Han, D.-S., Kim, H.-S., Jang, W.-H., Lee, S.-D., and Suh, J.-K., “Prespi: a domain combination based prediction system for protein–protein interaction,” *Nucleic acids research*, vol. 32, no. 21, pp. 6312–6320, 2004.
- [123] Jang, W.-H., Jung, S.-H., and Han, D.-S., “A computational model for predicting protein interactions based on multidomain collaboration,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1081–1090, 2012.
- [124] Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T., and Akiyama, Y., “Megadock: an all-to-all protein–protein interaction prediction system using tertiary structure data,” *Protein and peptide letters*, vol. 21, no. 8, pp. 766–778, 2014.

- [125] Pierce, B. and Weng, Z., “Zrank: reranking protein docking predictions with an optimized energy function,” *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 1078–1086, 2007.
- [126] Ohue, M., Matsuzaki, Y., Ishida, T., and Akiyama, Y., “Improvement of the protein–protein docking prediction by introducing a simple hydrophobic interaction model: An application to interaction pathway analysis,” in *Pattern Recognition in Bioinformatics*, pp. 178–187, Springer, 2012.
- [127] Matsuzaki, Y., Ohue, M., Uchikoga, N., and Akiyama, Y., “Protein-protein interaction network prediction by using rigid-body docking tools: Application to bacterial chemotaxis,” *Protein and peptide letters*, 2013.
- [128] Ohue, M., Matsuzaki, Y., Shimoda, T., Ishida, T., and Akiyama, Y., “Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods,” in *BMC Proceedings*, vol. 7, p. S6, BioMed Central Ltd, 2013.
- [129] Zhou, H., Pandit, S. B., and Skolnick, J., “Performance of the pro-sp3-tasser server in casp8,” *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 123–127, 2009.
- [130] Ishida, T. and Kinoshita, K., “Prediction of disordered regions in proteins based on the meta approach,” *Bioinformatics*, vol. 24, no. 11, pp. 1344–1348, 2008.
- [131] Deng, M., Mehta, S., Sun, F., and Chen, T., “Inferring domain–domain interactions from protein–protein interactions,” *Genome research*, vol. 12, no. 10, pp. 1540–1548, 2002.
- [132] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.*, “A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*,” *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.
- [133] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y., “Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 3, pp. 1143–1147, 2000.
- [134] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al.*, “The pfam protein families database,” *Nucleic acids research*, vol. 32, no. suppl 1, pp. D138–D141, 2004.
- [135] Schwikowski, B., Uetz, P., and Fields, S., “A network of protein–protein interactions in yeast,” *Nature biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.
- [136] Breiman, L., “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [137] Caruana, R., Karampatziakis, N., and Yessenalina, A., “An empirical evaluation of supervised learning in high dimensions,” in *Proceedings of the 25th international conference on Machine learning*, pp. 96–103, ACM, 2008.

- [138] Chang, K. Y. and Yang, J.-R., “Analysis and prediction of highly effective antiviral peptides based on random forests,” *PLoS ONE*, vol. 8, no. 8, p. e70166, 2013.
- [139] Izmirlian, G., “Application of the random forest classification algorithm to a seldi-tof proteomics study in the setting of a cancer prevention trial,” *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 154–174, 2004.
- [140] Qi, Y., “Random forest for bioinformatics,” in *Ensemble Machine Learning*, pp. 307–323, Springer, 2012.
- [141] Singh, R., Xu, J., and Berger, B., “Struct2net: Integrating structure into protein-protein interaction prediction,” in *Pacific Symposium on Biocomputing*, vol. 11, pp. 403–414, Citeseer, 2006.
- [142] Singh, R., Park, D., Xu, J., Hosur, R., and Berger, B., “Struct2net: a web service to predict protein–protein interactions using a structure-based approach,” *Nucleic acids research*, vol. 38, no. suppl 2, pp. W508–W515, 2010.
- [143] Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J., “Prediction of interactions between hiv-1 and human proteins by information integration,” in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 516, NIH Public Access, 2009.
- [144] Dyer, M. D., Murali, T., and Sobral, B. W., “Supervised learning and prediction of physical interactions between human and hiv proteins,” *Infection, Genetics and Evolution*, vol. 11, no. 5, pp. 917–923, 2011.
- [145] Mukhopadhyay, A., Ray, S., and Maulik, U., “Incorporating the type and direction information in predicting novel regulatory interactions between hiv-1 and human proteins using a biclustering approach,” *BMC bioinformatics*, vol. 15, no. 1, p. 26, 2014.
- [146] Zaki, N. and Campbell, P., “Domain linker region knowledge contributes to protein-protein interaction prediction,” in *Proceedings of International Conference on Machine Learning and Computing (ICMLC 2009)*, 2009.
- [147] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.*, “Pfam: the protein families database,” *Nucleic acids research*, p. gkt1223, 2013.
- [148] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., “Basic local alignment search tool,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.
- [149] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L., “Blast+: architecture and applications,” *BMC bioinformatics*, vol. 10, no. 1, p. 421, 2009.
- [150] Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R., “Domine: a database of protein domain interactions,” *Nucleic acids research*, vol. 36, no. suppl 1, pp. D656–D661, 2008.
- [151] Kim, Y., Min, B., and Yi, G.-S., “Iddi: integrated domain-domain interaction and protein interaction analysis system,” *Proteome Sci*, vol. 10, no. Suppl 1, p. S9, 2012.

- [152] Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P., “3did: a catalog of domain-based interactions of known three-dimensional structure,” *Nucleic acids research*, p. gkt887, 2013.
- [153] Sasaki, Y., “The truth of the f-measure,” *Teach Tutor mater*, pp. 1–5, 2007.
- [154] Powers, D., “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [155] Hanley, J. A. and McNeil, B. J., “A method of comparing the areas under receiver operating characteristic curves derived from the same cases.,” *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.
- [156] Zaki, N., Bouktif, S., and Lazarova-Molnar, S., “A combination of compositional index and genetic algorithm for predicting transmembrane helical segments,” *PloS ONE*, vol. 6, no. 7, p. e21821, 2011.
- [157] Kirkpatrick, S., Jr., D. G., and Vecchi, M. P., “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [158] Vega-Rodríguez, M. A., Gómez-Pulido, J. A., Alba, E., Vega-Pérez, D., Priem-Mendes, S., and Molina, G., “Evaluation of different metaheuristics solving the rnd problem,” in *Applications of Evolutionary Computing*, pp. 101–110, Springer, 2007.
- [159] Mendes, S. P., Molina, G., Vega-Rodríguez, M. A., Gómez-Pulido, J. A., Sáez, Y., Miranda, G., Segura, C., Alba, E., Isasi, P., León, C., *et al.*, “Benchmarking a wide spectrum of metaheuristic techniques for the radio network design problem,” *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 5, pp. 1133–1150, 2009.
- [160] Busetti, F., “Simulated annealing overview,”
- [161] Henderson, D., Jacobson, S. H., and Johnson, A. W., “The theory and practice of simulated annealing,” in *Handbook of metaheuristics*, pp. 287–319, Springer, 2003.
- [162] Geman, S. and Geman, D., “Stochastic relaxation, gibbs distributions, and the bayesian restoration of images,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.
- [163] Salamon, P., Sibani, P., and Frost, R., *Facts, conjectures, and improvements for simulated annealing*.  
SIAM, 2002.
- [164] Ingber, L., “Simulated annealing: Practice versus theory,” *Mathematical and computer modelling*, vol. 18, no. 11, pp. 29–57, 1993.
- [165] Smith, K. I., Everson, R. M., Fieldsend, J. E., Murphy, C., and Misra, R., “Dominance-based multiobjective simulated annealing,” *Evolutionary Computation, IEEE Transactions on*, vol. 12, no. 3, pp. 323–342, 2008.
- [166] Hansen, P. B., “Simulated annealing,”
- [167] Jaraíz-Simon, M. D., Gómez-Pulido, J. A., Vega-Rodríguez, M. A., and Sánchez-Pérez, J. M., “Simulated annealing for real-time vertical-handoff in wireless networks,” in *Advances in Computational Intelligence*, pp. 198–209, Springer, 2013.

- [168] Kundu, S., Mahato, M., Mahanty, B., and Acharyya, S., “Comparative performance of simulated annealing and genetic algorithm in solving nurse scheduling problem,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2008.
- [169] Bertsimas, D. and Nohadani, O., “Robust optimization with simulated annealing,” *Journal of Global Optimization*, vol. 48, no. 2, pp. 323–334, 2010.
- [170] Huang, K.-Y. and Hsieh, Y.-H., “Very fast simulated annealing for pattern detection and seismic applications,” in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, pp. 499–502, IEEE, 2011.
- [171] Ho, S.-Y., Hsieh, C.-H., Yu, F.-C., and Huang, H.-L., “An intelligent two-stage evolutionary algorithm for dynamic pathway identification from gene expression profiles,” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 4, no. 4, pp. 648–704, 2007.
- [172] Noel, V., Vakulenko, S., and Radulescu, O., “Algorithm for identification of piecewise smooth hybrid systems: application to eukaryotic cell cycle regulation,” in *Algorithms in Bioinformatics*, pp. 225–236, Springer, 2011.
- [173] Tomshine, J. and Kaznessis, Y. N., “Optimization of a stochastically simulated gene network model via simulated annealing,” *Biophysical journal*, vol. 91, no. 9, pp. 3196–3205, 2006.
- [174] Bryan, K., Cunningham, P., and Bolshakova, N., “Biclustering of expression data using simulated annealing,” in *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, pp. 383–388, IEEE, 2005.
- [175] Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A., and Kanehisa, M., “Multiple sequence alignment by parallel simulated annealing,” *Computer applications in the biosciences: CABIOS*, vol. 9, no. 3, pp. 267–273, 1993.
- [176] Kim, J., Pramanik, S., and Chung, M. J., “Multiple sequence alignment using simulated annealing,” *Computer applications in the biosciences: CABIOS*, vol. 10, no. 4, pp. 419–426, 1994.
- [177] Frith, M. C., Hansen, U., Spouge, J. L., and Weng, Z., “Finding functional sequence elements by multiple local alignment,” *Nucleic acids research*, vol. 32, no. 1, pp. 189–200, 2004.
- [178] Shatnawi, M. and Zaki, N., “Prediction of protein inter-domain linkers using compositional index and simulated annealing,” in *Proceeding of the fifteenth annual conference companion on Genetic and evolutionary computation conference companion*, GECCO '13 Companion, pp. 1603–1608, 2013.
- [179] Hu, H.-J., Pan, Y., Harrison, R., and Tai, P. C., “Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier,” *NanoBioscience, IEEE Transactions on*, vol. 3, no. 4, pp. 265–271, 2004.
- [180] Kim, H. and Park, H., “Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3d local descriptor,” *Proteins: Structure, Function, and Bioinformatics*, vol. 54, no. 3, pp. 557–562, 2004.
- [181] Korenberg, M. J., David, R., Hunter, I. W., and Solomon, J. E., “Automatic classification of protein sequences into structure/function groups via parallel

- cascade identification: a feasibility study,” *Annals of biomedical engineering*, vol. 28, no. 7, pp. 803–811, 2000.
- [182] Yoo, P., Zhou, B., and Zomaya, A., “A modular kernel approach for integrative analysis of protein domain boundaries,” *BMC genomics*, vol. 10, no. Suppl 3, p. S21, 2009.
- [183] David, R., *Applications of nonlinear system identification to protein structural prediction*.  
PhD dissertation, Massachusetts Institute of Technology, 2000.
- [184] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H., “Hydrophobicity of amino acid residues in globular proteins,” *Science*, vol. 229, no. 4716, pp. 834–838, 1985.
- [185] Taylor, W. R., “The classification of amino acid conservation,” *Journal of theoretical Biology*, vol. 119, no. 2, pp. 205–218, 1986.
- [186] Betts, M. J. and Russell, R. B., “Amino acid properties and consequences of substitutions,” *Bioinformatics for geneticists*, vol. 317, p. 289, 2003.
- [187] Ganapathiraju, M., Balakrishnan, N., Reddy, R., and Klein-Seetharaman, J., “Transmembrane helix prediction using amino acid property features and latent semantic analysis,” *Bmc Bioinformatics*, vol. 9, no. Suppl 1, p. S4, 2008.
- [188] Hayat, M. and Khan, A., “Mem-phybrid: Hybrid features-based prediction system for classifying membrane protein types,” *Analytical biochemistry*, vol. 424, no. 1, pp. 35–44, 2012.
- [189] Hayat, M. and Khan, A., “Wrf-tmh: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids,” *Amino acids*, pp. 1–12, 2013.
- [190] Wang, X.-F., Chen, Z., Wang, C., Yan, R.-X., Zhang, Z., and Song, J., “Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach,” *PLoS ONE*, vol. 6, no. 10, p. e26767, 2011.
- [191] Yang, P., Hwa Yang, Y., B Zhou, B., and Y Zomaya, A., “A review of ensemble methods in bioinformatics,” *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.
- [192] Lee, J. W., Lee, J. B., Park, M., and Song, S. H., “An extensive comparison of recent classification tools applied to microarray data,” *Computational Statistics & Data Analysis*, vol. 48, no. 4, pp. 869–885, 2005.
- [193] Díaz-Urriarte, R. and De Andres, S. A., “Gene selection and classification of microarray data using random forest,” *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.
- [194] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H., “Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data,” *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.
- [195] Chen, C., Liaw, A., and Breiman, L., “Using random forest to learn imbalanced data,” *University of California, Berkeley*, 2004.

- [196] Bordner, A. J., “Predicting protein-protein binding sites in membrane proteins,” *BMC bioinformatics*, vol. 10, no. 1, p. 312, 2009.
- [197] Šikić, M., Tomić, S., and Vlahoviček, K., “Prediction of protein-protein interaction sites in sequences and 3d structures by random forests,” *PLoS computational biology*, vol. 5, no. 1, p. e1000278, 2009.
- [198] Han, P., Zhang, X., Norton, R., and Feng, Z.-P., “Large-scale prediction of long disordered regions in proteins using random forests,” *BMC bioinformatics*, vol. 10, no. 1, p. 8, 2009.
- [199] Wang, C., Xi, L., Li, S., Liu, H., and Yao, X., “A sequence-based computational model for the prediction of the solvent accessible surface area for  $\alpha$ -helix and  $\beta$ -barrel transmembrane residues,” *Journal of computational chemistry*, vol. 33, no. 1, pp. 11–17, 2012.
- [200] Hernández-Lobato, D., Martínez-Muñoz, G., and Suárez, A., “How large should ensembles of classifiers be?,” *Pattern Recognition*, vol. 46, no. 5, pp. 1323–1336, 2013.
- [201] Bibimoune, M., Elghazel, H., and Aussem, A., “An empirical comparison of supervised ensemble learning approaches,” in *International Workshop on Complex Machine Learning Problems with Ensemble Methods COPEM@ ECML/PKDD*, vol. 13, pp. 123–138, 2013.
- [202] Williams, A. F. and Barclay, A. N., “The immunoglobulin superfamily-domains for cell surface recognition,” *Annual review of immunology*, vol. 6, no. 1, pp. 381–405, 1988.
- [203] Rubinstein, E., Ziyat, A., Wolf, J.-P., Le Naour, F., and Boucheix, C., “The molecular players of sperm-egg fusion in mammals,” in *Seminars in cell & developmental biology*, vol. 17, pp. 254–263, Elsevier, 2006.
- [204] Inoue, N., Ikawa, M., Isotani, A., and Okabe, M., “The immunoglobulin superfamily protein izumo is required for sperm to fuse with eggs,” *Nature*, vol. 434, no. 7030, pp. 234–238, 2005.
- [205] Needleman, S. B. and Wunsch, C. D., “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.
- [206] Chan, A., “An analysis of pairwise sequence alignment algorithm complexities: Needleman-wunsch, smith-waterman, fasta, blast and gapped blast,” 2007.
- [207] Zvelebil, M. and Baum, J., *Understanding bioinformatics*. Garland Science, 2007.

## List of Publications

The following is a list of the published articles related to this dissertation:

### Journal Articles:

- Maad Shatnawi and Nazar Zaki (2015) Inter-domain linker prediction using amino acid compositional index. *Computational Biology and Chemistry (CBAC)* 55: 23-30, April 2015. (ISI IF 1.595)
- Maad Shatnawi, Nazar Zaki, and Paul D. Yoo (2014) "Protein inter-domain linker prediction using random forest and amino acid physiochemical properties." *BMC Bioinformatics* 15 (Suppl 16): S8, December 2014. (ISI IF 2.670)

### Conference Papers:

- Maad Shatnawi and Nazar Zaki (2013) Prediction of protein inter-domain linkers using compositional index and simulated annealing. In *Proceeding of the ACM 15th annual conference on Genetic and evolutionary computation (GECCO '13)*, pp: 1603-1608, Amsterdam, Netherland, July 2013. (Ranked A)
- Maad Shatnawi and Nazar Zaki (2013) Prediction of inter-domain linker regions in protein sequences: A survey. In the *proceeding of the 9th International Conference on Innovations in Information Technology (IIT)*, pp: 237-242, Al-Ain, UAE, March 2013.
- Maad Shatnawi (2014) "Computational Methods for Protein-Protein Interaction Prediction". In *Proceedings of the International Conference on Bioinformatics and Computational Biology (BIOCOMP'14)*, Las Vegas, USA, July 2014.
- Maad Shatnawi and Nazar Zaki Novel domain identification approach for protein-protein interaction prediction. *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, Niagara Falls, Canada, August 2015.

### Book Chapters:

- Maad Shatnawi, Paul D. Yoo, and Sami Muhaidat (2015) Protein inter-domain linker prediction. In "Pattern Recognition in Computational Molecular Biology: Techniques and Approaches," *Wiley Series in Bioinformatics*, August 2015, Wiley.
- Maad Shatnawi (2015) Review of the recent protein-protein interaction techniques. In "Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology - Algorithms and Software Tools". July 2015, Elsevier/MK.