DETECTING MOTIFS FOR COMPUTATIONAL CLASSIFICATION OF DOCKERIN AND COHESIN SEQUENCES

by
Ebru Şahin
2013

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
January 2013

DETECTING MOTIFS FOR COMPUTATIONAL CLASSIFICATION OF DOCKERIN AND COHESIN SEQUENCES

APPROVED BY:

Prof. Dr. Osman Uğur Sezerman .....
(Thesis Advisor)

Assoc. Prof. Devrim Gözüaçık .....

Assoc. Prof. Tonguç Ünlüyurt .....

Assist. Prof. Kemal Kılıç .....

Assoc. Prof. Levent Öztürk .....

DATE OF APPROVAL: 28.01.2013

to my family

&

my fiancée ihsan

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Osman Uğur Sezerman for his continuous support and encouragement throughout this thesis. I am very thankful to my thesis committee members Levent Öztürk, Devrim Gözüaçık, Tonguç Ünlüyurt and Kemal Kılıç for their valuable comments and suggestions on this thesis.

I would like to thank to İhsan Kehribar for his technical and moral support.

I am indebted to Tübitak for providing financial support during my studies.

I would like to express my special appreciation to my family for their unconditional love and support.

DETECTING MOTIFS FOR COMPUTATIONAL CLASSIFICATION OF DOCKERIN AND COHESIN SEQUENCES

Ebru Şahin

Ms Thesis, 2013

Thesis Supervisor: Prof. Dr. Osman Uğur Sezerman

Keywords: Cellulosome, Dockerin Classification, Cohesin Classification, Motif Detection, Reduced Amino Acid Alphabets, Correlated Mutation.

ABSTRACT

Cellulose is the most abundant biopolymer in nature. It has several usage areas in industry. The initial hydrolysis of cellulose is the rate determining step in cellulose degradation. Cellulosomes are the complex structures composed of non-catalytic units and enzymes that take part in cellulose degradation. Cellulosomal units are attached via the interaction between cohesin and dockerin domains which are divided into three subclasses; type I, type II and type III. Development and rational design of novel cohesin and dockerin domains to enhance synergistic actions is very important research topic for biotechnological applications. In this aspect, accurate classification of the subunits and identification of key interaction sites are of great importance for design purposes.

In this thesis, we propose a multiple sequence alignment and information theory based classification method that identifies potential key interaction sites. Based on the multiple sequence alignments, the residues that are conserved only in one subclass are determined as the motifs. Classification performance of these motifs is determined using a majority voting based normalized scoring scheme. In addition, reduced amino acid alphabets are introduced to capture the similarities that are invisible in 20-letter alphabet.

In this work, we classify cohesin sequences with 99% accuracy, 96% sensitivity and 97% specificity, on average. For dockerin, the sequences are classified with up to 95% accuracy. 76% sensitivity and 92% specificity are observed on average. Potential interaction sites between cohesins and dockerins are determined from the correlated mutation analysis.

DOCKERİN VE KOHEZİN DİZİLERİNİN HESAPLAMALI SINIFLANDIRILMASI İÇİN MOTİFLERİN TESPİTİ

Ebru Şahin

Ms Tezi, 2013

Tez Danışmanı: Prof. Dr. Osman Uğur Sezerman

Anahtar Kelimeler: Selülozom, Dockerin Sınıflandırılması, Kohezin Sınıflandırılması, Motif Tespiti, İndirgenmiş Aminoasit Alfabeleri, İlintili Mutasyon.

Özet

Selüloz doğada en yaygın bulunan biyopolimerdir. Selülozun sanayide çok çeşitli kullanım alanları mevcuttur. Selülozun ilk hidrolizi, selüloz yıkımındaki hız beliryici basamaktır. Selülozom, katalitik olmayan birimlerden ve selüloz yıkımında rol alan enzimlerden oluşan kompleks bir yapıdır. Selülozomun yapısal birimleri birbirlerine kohezin ve dockerin bölgeleri arasındaki etkileşim ile bağlanır. Dockerin ve kohezin bölgeleri tip I, tip II ve tip III olmak üzere üç alt gruba ayrılır. Enzimler arasındaki sinerjik işleyişin artırılması amacıyla yeni kohezin ve dockerin bölgelerinin dizaynı ve geliştirilmesi biyoteknoloji uygulamaları için önemli araştırma konularından biridir. Bu çerçevede, dockerin ve kohezin alt gruplarının doğru bir biçimde sınıflandırılması ve anahtar etkileşim noktalarının tanımlanması dizayn çalışmaları için büyük önem arzetmektedir.

Bu çalışmada, çoklu dizi hizalaması temelli ve potansiyel anahtar etkileşim noktalarını açığa çıkaran bir sınıflandırma metodu tanıtıyoruz. Çoklu dizi hizalamalarını kullanarak, yalnızca bir alt grupta korunmuş aminoasitler ve lokasyonları motif olarak tanımlandı. Motiflere ait sınıflandırma performansları, çoğunluk oylaması temelli normalize edilmiş bir skor şeması kullanılarak belirlendi. Ayrıca, 20-harfli aminoasit alfabesinde görünmeyen benzerlikleri yakalamak için indirgenmiş aminoasit alfabeleri tanıtıldı.

Bu çalışmada, kohezin dizileri %99'e varan oranda doğru sınıflandırıldı. Ayrıca, ortalama %96 hassasiyet ve %97 spesifiklik elde edildi. Dockerin dizileri %95'e varan oranda doğru sınıflandırılırken, ortalama %76 hassasiyet ve % 92 spesifiklik elde edildi. Potansiyel anahtar etkileşim noktaları ilintili mutasyon analizi kullanılarak tanımlandı.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF ABBREVIATIONS

CBD	Carbohydrate Binding Domain
PDB	Protein Data Bank
RAAA	Reduced Amino Acid Alphabet
HMM	Hidden Markov Model
SVM	Support Vector Machine
MSA	Multiple Sequence Alignment
PS	Presence in Subclass
SS	Subclass Specificity
MSS	Motif Specificity Score
CS	Classification Score
MEP	Minimum Error Point
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

Chapter 1

INTRODUCTION

1.1 Motivation

Cellulose, a major component of the plant cellwalls, is the most abundant biopolymer in nature. Cellulose is constructed into a tightly packed and highly ordered structure, through extensive hydrogen bonding and van der Waals stacking interactions. Packed and ordered structure of cellulose, as well as its association with other structural polymers make the cellulose considerably resistant to microbial degradation [1, 2].

Cellulose as the most abundant biopolymer on Earth is additionally the most abundant renewable carbon and energy source in nature. Consequently, degradation of cellulose to smaller carbon compounds is an essential process for carbon cycle in nature [3]. In addition to its importance for nature, smaller carbon compounds gained considerable attention as alternative, environment friendly energy source [4]. In the modern age, biorefineries are being developed as a clean alternative to the fossil fuels and cellulose degradation appears as a fundamental process to produce smaller carbon sources to be consumed in these biorefineries [5]. Besides, cellulosic compounds have an excessive potential to be benefited for several products in biotechnology based industries and for food applications [6]. In order to utilize this potential, several studies are being conducted on the initial hydrolysis of cellulose, the rate-determining step for cellulose utilization. For that purpose, cellulose degrading enzymes, their complexes and their working mechanism is an attractive research object [7].

Cellulosome is an extra-cellular, large supramolecular complex that have been identified in several bacteria [8]. Enzymes that take a part in cellulose degradation (e.g cellulases, hemi-cellulases) are assembled into cellulosomes with numerous other non-catalytic integrating proteins, called scaffoldin. Scaffoldins interact with cellulosomal enzymes through their cohesin domain [9]. The dockerin domains from enzymes interact tightly with cohesins. In some bacteria, several scaffoldins form a complex in cellulosome, and their attachment to each other is also secured through dockerin-cohesin interactions. Cohesin and dockerin domains are divided into three distinguished classes: type I, type II and type III. The interactions between cohesin-dockerin domains are type specific, exhibiting no cross-reactivity [10].

As stated above, cellulosomal subunits attract attention of scientists due to environmental problems, useful applications in industry and capacity of cellulose as an energy source. For example, designer cellulosome concept, the artificial enzymatic complex with increased degradation efficiency, is one of the hot topics in this area [11]. In this context, the efficiency of the complex is targeted by several different approaches. Artificial addition of cohesin and dockerin subunits to enzymes or scaffoldins to recruit enzymes of interest into cellulosome complex is applied several times, for different enzymes and different cohesin-dockerin interaction types [12]. In addition to the incorporation of enzymes into the cellulosome; development of novel cohesin and dockerin domains, and rational design or directed mutagenesis of cellulosomal components to enhance synergistic actions are hot research topics in designer cellulosome development [13]. At this point, accurate classification of the subunits and identification of key interaction sites gain considerable importance.

Analysis of dockerin-cohesin interactions holds key for both scientific and technological purposes. The origins of the specificity between subclasses of cohesins and dockerins are still not clearly understood and this is a significant scientific interest in order to fully comprehend the cellulosome organizations. The limited structures of cohesin-dockerin complexes provide an image, however this information does not reveal adequate information to design novel cohesin-dockerin interactions [14-17]. At this juncture, classification into subclasses (type I, type II and type III) and understanding of class specific key factors that governs the highly-specific dockerin-cohesin interaction appears to be a key challenge.

In this thesis, we propose a multiple sequence alignment and information theory based method for classification of dockerin and cohesin sequences. On the contrary of other computational approaches, this method allows identifying informative amino-acid residues in classes that are important for class specificity and also for their function; which comprise key-site candidates for interaction sites. In our method, the sequences including in type I, type II and type III classes are aligned separately. Working on the consensus sequences, the amino-acids conserved at a certain residue in one class but not in any other, are defined as motifs and given scores based on their specificity. Those motifs are then used to make classifications, calculating scores for test sequences. Utilization of Reduced Amino acid Alphabets to identify physiochemical conserved amino acids increases the accuracy of classification for several other protein families, eliminating the errors caused by incompetence of multiple alignments. In addition, RAAAs facilitate the identification of physiochemical properties for cohesin-dockerin families that are important for family specific cohesin- dockerin interaction; thus enabling the understanding of the mechanism of interaction. In this study, four different reduced amino-acid alphabets are introduced, in order to explore the effects of these RAAAs on the accuracy of classification. Subsequently, an HMM-based classification is carried out to compare out approach with a state-of-the-art classification method. Lastly, to identify key interaction sites between cohesions and dockerins for design purposes, we carry out correlated mutation studies in order to affirm the biological importance of detected key site candidates.

1.2 Outline

The organization of thesis as follows: Chapter 2 gives a brief biological background and an overview of computational methods that is used for protein classification. Methods that are used in this study are explained in detail, in Chapter 3. In Chapter 4, the results of the classification of cohesin and dockerin families along with the correlated mutation studies are presented. Lastly, Chapter 5 summarizes the conclusions and discusses future works.

Chapter 2

BACKGROUND AND RELATED WORKS

2.1 Biological Background

2.1.1 Proteins, Structure and Function

A protein is composed of amino acids that are attached together by peptide bonds. In nature, there are 20 amino acids with distinct biochemical properties, such as polar, hydrophobic and charge characters. Amino acids are constituted by an amino group (-NH_2), a carboxyl group (-COOH), a side chain and a central carbon atom adhered to the mentioned groups. Except for side chains, the other components of the amino acids occur to be the same. Side chains, on the other hand, are the components that contribute to the distinct biochemical properties of amino acids [18].

During the protein synthesis, carboxyl group of one amino acid and amino group of another form a peptide bond, producing a water molecule [19]. The amino acids joined together via peptide bonds form the primary structure of proteins. Concurrently, hydrogen bonds constructed between backbone atoms contribute to the formation of secondary structure elements, such as alpha (α) helices and beta (β) sheets [20].

Following the secondary structure formation, the attractions between α-helices and β-sheets arising from the side chains form a spatial arrangement. The peptide chain is folded into a 3-dimensional, biologically active state, named tertiary structure. Functionally fundamental parts of proteins such as catalytic sites and binding sites are

formed by tertiary structures. Therefore, accurate folding of proteins into their 3D structure is of basic importance for their function [21].

As stated above briefly, the interactions that induce 3-D folding are emanating from biochemical properties or amino acid side chains. H-bonds, van der Waals interactions, backbone angle preferences, electrostatic and hydrophobic interactions drive the protein into its 3-D functional structure. These interactions between amino acids are controlled by their side chain structure and properties, such as hydrophobicity, polarity or charges. Therefore, the distribution of hydrophobic and hydrophilic residues in a protein has great impact on the total tertiary structure of the protein [22]. Amino acids and their basic biochemical properties are summarized in Table 2.1 [18].

Table 2.1 List of amino acids and their biochemical properties [18].

Amino Acid	Abbreviations		Hydropathy Index	Polarity	Charge
	3 letter	Single Letter			
Isoleucine	Ile	I	4.5	nonpolar	neutral
Valine	Val	V	4.2	nonpolar	neutral
Leucine	Leu	L	3.8	nonpolar	neutral
Phenylalanine	Phe	F	2.8	nonpolar	neutral
Cysteine	Cys	C	2.5	nonpolar	neutral
Methionine	Met	M	1.9	nonpolar	neutral
Alanine	Ala	A	1.8	nonpolar	neutral
Glycine	Gly	G	-0.4	nonpolar	neutral
Threonine	Thr	T	-0.7	polar	neutral
Tryptophan	Trp	W	-0.9	nonpolar	neutral
Serine	Ser	S	-0.8	polar	neutral
Tyrosine	Tyr	Y	-1.3	polar	neutral
Proline	Pro	P	-1.6	nonpolar	neutral
Histidine	His	H	-3.2	polar	positive
Glutamic acid	Glu	E	-3.5	polar	negative
Glutamine	Gln	Q	-3.5	polar	neutral
Aspartic acid	Asp	D	-3.5	polar	negative
Asparagine	Asn	N	-3.5	polar	neutral
Lysine	Lys	K	-3.9	polar	positive
Arginine	Arg	R	-4.5	polar	positive

2.1.2 Cellulose as a Structural Component

Cellulose, a major component of the plant cellwalls, is the most abundant biopolymer in nature. Plant cellwalls are reinforced by the cross-linked structure of cellulose microfibrils, whose insoluble nature is ideal to secure structural stability [1, 23].

The backbone structure of cellulose is consisted of unbranched (1,4) β-linked D-glucose [24]. Adjacent D-glucose units are flipped forming cellobiose, the structural repetitive unit of cellulose (Figure 2.1). Linear cellulose polymer exhibits a dense intermolecular bonding pattern. Accordingly, cellulose chains are tightly packed and organized in parallel generating crystalline microfibrils [25]. Despite its bare chemical composition, microfibrils do incorporate less ordered, non-crystalline regions, as well as highly ordered crystalline region. Those amorphous parts are more susceptible to enzymatic degradation and generally featured on cellulose surface[26] [27].

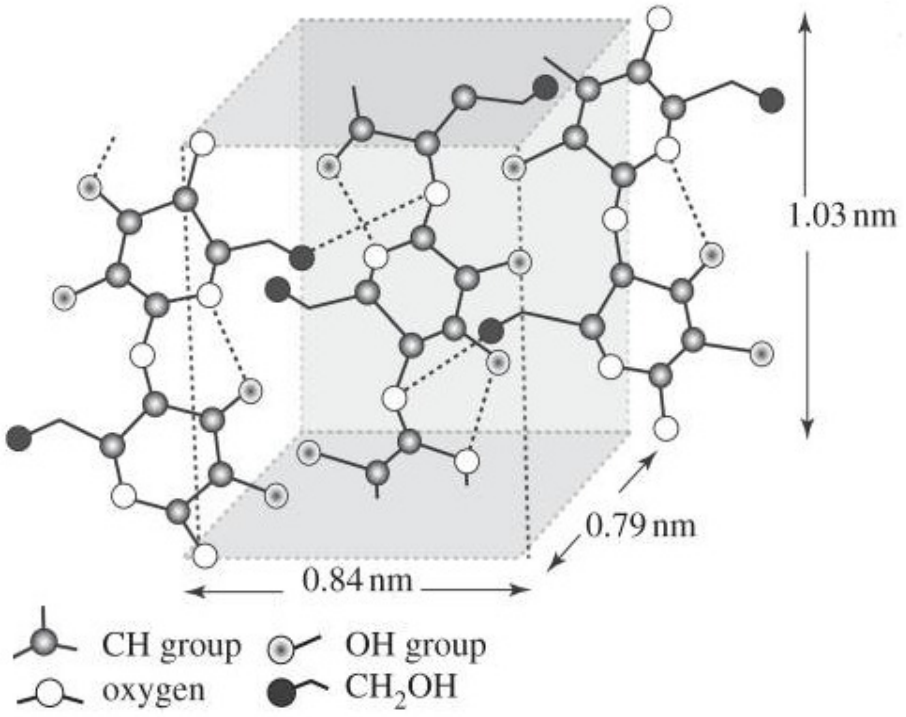

Figure 2.1 The structure and the inter- and intra-chain hydrogen bonding pattern in cellulose [8].

Through extensive hydrogen bonding and van der Waals stacking interactions, microfibrils are able to form non-covalent complexes which leads to tightly packed macrofibrillar structures [2] [1]. Those macrofibillar structures of cellulose are aligned by a matrix of hemicellulose and either lignin or pectin polysaccharides in cell wall construction. The volume fraction of these building blocks can vary based on the specie, tissue type and differing growth patterns [28].

Tightly packed and highly ordered construction of cellulose, its association with other structural polymers and its insoluble nature makes the cellulose considerably resistant to microbial degradation. Although cellulose is formed by a single type of chemical bond and has a chemically simplistic structure, multiple enzyme systems are required for effective degradation [1, 25].

2.1.2.1 Importance of Cellulose Degradation

Cellulose, the most abundant biopolymer on Earth, is additionally the most abundant renewable carbon and energy source in nature, with 180 million tons raw material capacity per year. Consequently, degradation of cellulosic biomass is an essential process for carbon cycle and arousing interest as a bio-energy source [5].

Carbon cycle, in general, can be summarized as fixation of carbon through photosynthesis and formation of CO_2 from those fixated carbon sources through combustion [3]. In order to metabolize cellulose to CO_2, the crystalline cellulose has to be degraded enzymatically to yield cellobiose and then, converted to glucose by β-glucosidase [29]. Cellulose, as a major carbon source and its recycling by microorganisms are therefore imperative in the carbon cycle [3, 4].

In the modern age; as the fossil fuels will be exhausted in the near future and the earth is facing serious environmental problems like global warming; new alternative and environment friendly energy sources has gained considerable importance [4]. Thence, biorefineries are being developed to use bio-fuel as an alternative energy source and consequently, cellulose degradation appears as a fundamental process to produce smaller carbon sources to be consumed in those biorefineries [5, 30].

Plants are being used widely in industrial fields to produce furniture, paints, fabrics, medicine, paper, food, bio-ethanol and several other products, yielding a cellulosic bio-

mass as waste [30-32]. The accumulation of cellulosic waste arises as an environmental problem. However, more to the point, the cellulosic products labeled as "waste", has an excessive potential to be benefited for recovery of several products in biotechnology based industries and for food applications [6]. In order to utilize this potential, several researches are being conducted on microorganisms which can process cellulosic compounds. For those microorganisms, the ability to degrade cellulose compounds to smaller sugars effectively with minimum pre-processing is an important feature, forasmuch as, the mentioned applications mostly requires hydrolysis of cellulose initially [7]. On the ground that the initial hydrolysis of cellulose is the rate-determining step for cellulose utilization; cellulose degrading enzymes, their complexes and their working mechanism become an attractive research object [5].

2.1.3 The Cellulosome Complex

Several bacteria and fungi produce a variety of enzymes, called cellulases that catalyze degradation of crystalline cellulose, and thus, plant cellwalls. Heterogeneous, insoluble and recalcitrant nature of plant cellwalls complicates the process of degradation, even though a single type of chemical bond is being targeted by enzymes. For years, it is thought that several free cellulases work synergistically on that complex nature of crystalline cellulose, creating an enzyme system. Although this case is true for many aerobic microorganisms, the discovery of cellulosome complex broadened the knowledge about cellulase enzyme systems [8, 33].

In aerobes, numerous cellulase enzymes are either secreted to extracellular matrix or bound to the outer membrane. Even though the enzymes are not physically adhered, they act in strong synergy to degrade complex, crystalline cell wall cellulose [23]. In anaerobic microorganisms, however, cellulase enzymes are assembled into large, supramolecular, surface-attached structures, called cellulosome. In cellulosome complex, a variety of cellulases and hemi-cellulases are tightly adhered to a central, multimodular, non-catalytic integrating protein, called *scaffoldin [9, 17]*. Scaffoldins interact with cellulosomal enzymes through a specific domain, named *cohesin*. Scaffoldins contain numerous cohesin domains that interact with another specific type of domain from cellulosomal enzymes, named *dockerin*. The cohesin-dockerin interaction is the funda-

mental molecular mechanisms that secures the integration of enzymes into the cellulosome complex [10].

It is widely believed that the major function of cellulosome is to bring cellulases into close proximity to potentiate synergy between different catalytic components [34]. On the other hand, the synergy may be reduced due to conformational restrictions emanated from the physical association of enzymes within the complex structure of cellulosome. In order to address that question, several studies demonstrate that cellulosome ensemble has crucial conformational flexibility and congregating the enzymes induces approximately threefold increase in synergy [33].

Cellulosome complex does not merely gather catalytic components to increase synergy, but also locates the enzymes in the vicinity of cellulosic compounds. Exhibition of enzyme complex on the cell surface is a remarkable feature that facilitates the efficient consumption of cellulosic products by microorganisms [34]. Additionally, scaffoldins possess a cellulose-specific carbohydrate-binding domain (CBD) for substrate targeting. In different species, however, the mechanism of carbohydrate-binding can show variations, such as a necessity of additional scaffoldins.

2.1.3.1 Cellulosome Associated Elements

Bacterial cellulosome systems display diversity among different species. Mainly two different cellulosome systems are differentiated, as simple and complex cellulosome systems [25]. In simple cellulosome systems, scaffoldins own a single CBD, several cohesin domains and one or more X modules, with unknown function. Dockerin-borne cellulosomal enzymes interact with the cohesin domains of scaffoldin and attached to the cellulosome complex [35, 36]. Scaffoldins in simple cellulosome systems are associated with the cell surface; however, the exact molecular mechanism is unclear. Those types of scaffoldins are named as *primary scaffoldins [37]*.

Complex cellulosome systems, on the other hand; contain several scaffoldins that are attached to each other in different ways, constituting the complex form of the cellulosome. In those systems, one of the scaffoldins functions as a primary scaffoldin and recruit dockerin-borne cellulosomal enzymes into the complex. However, in contrast to the primary scaffoldins in simple cellulosome systems, those scaffoldins contain a dif-

ferent type of dockerin subunit, in addition to its cohesin subunits [38, 39]. In order to tether the cellulosome complex to the cell surface, the additional dockerin subunit interacts with cohesins from other scaffoldins. The scaffoldins that incorporates the cellulosome to the cell membrane are named as *anchoring scaffoldins [40]*. Moreover, various complex cellulosomes involve additional scaffoldins that enhance the number of components in cellulosome, named *adaptor scaffoldins [41]*.

2.1.3.2 Dockerin and Cohesin Subunits in Cellulosomes

As mentioned above, the cohesin-dockerin interaction is the fundamental key for the assembly of cellulosome complex. In primary scaffoldins, cohesins exist as highly homologous repetitive units that dock the cellulosomal enzymes to the complex cellulosome [17]. Enzymes interact with scaffoldin through their dockerin domain. The existence of dockerin subunit is the major difference that distinguishes cellulosomal enzymes from non-cellulosomal ones [25].

Additional anchoring or adaptor scaffoldins involved in the cellulosome are attached to the primary scaffoldin through cohesin-dockerin interactions [9]. However, the cohesin domains in those additional scaffoldins display a different character and do not interact with the dockerin domains from enzymes [10]. In this context, known cohesin and dockerin sequences are identified in three distinguished subgroups: type I, type II and type III. Several interaction studies demonstrate that cohesins and dockerins belonging in one subgroup only interact with dockerins and cohesins in that specific group. Put another way, there is no observed cross-reactivity between type I, type II and type III elements [35].

2.1.3.2.1 Type I cohesin-dockerin Interaction

The mechanisms of type I cohesin-dockerin interaction is revealed by several structural studies. The individual structures of dockerin and cohesins are also studied and provide noteworthy information about the interaction process. In 1997, Shimon et al. defined type I cohesin modules by a jelly roll topology composed of nine β-strands fold in two β-sheets [42].

Shortly after cohesin type I structure is determined, in 2001, Lytle et al. revealed the solution structure of type I dockerin domain, which is composed of three α-helices [17] . In detail, type I dockerin contains tandem duplication of a 22-residue sequence and helices at the N-terminal and C-terminal ends are formed by this 22-residue repeats, in addition F-hand type calcium-binding motifs [43]. The structural conservation among the repeated segments is remarkable and thus, the N-terminal duplicated segment can be superimposed over the C-terminal duplicated segment, providing the structural basis for the dual mode of binding [44].

Structural data of dockerin-cohesin in complex demonstrates that type I dockerins display two identical cohesin binding interfaces. Dockerin could be rotated 180° relative to its initial position, therefore; in one mode N-terminal helix (helix 1) concludes cohesin recognition and in the second binding mode, dockerin is flipped 180° relative to the cohesin and C-terminal helix (helix 3) dominates the ligand recognition [45]. In addition, presence of Ca^{+2} is essential for dockerin-cohesin interaction [45].

(a) (b)

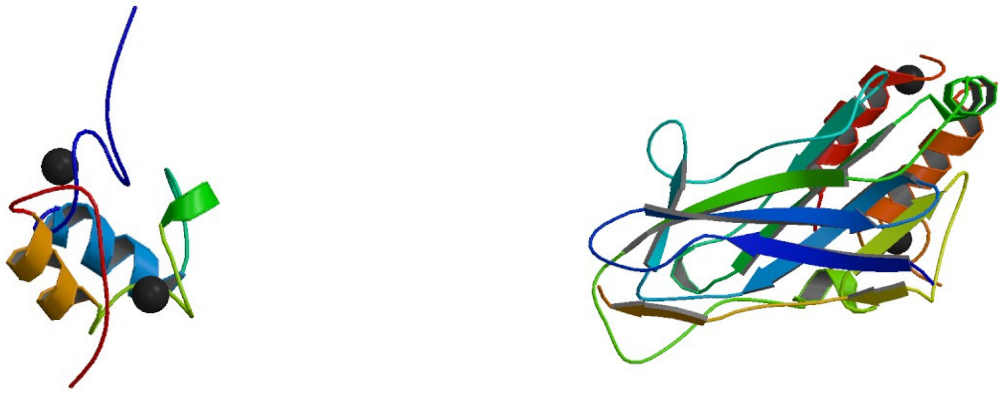

Figure 2.2 (a) Internal symmetry of WT type I dockerin in complex with two Ca^{+2} ions from Clostridium thermocellum cellulosome (PDB code: 1 DAQ) (b) Type II cohesin-dockerin interaction from Bacteroides cellolosolvens (PDB code: 2Y3N)

2.1.3.2.2 Type II cohesin-Dockerin Interaction

Structural studies on type II cohesin-dockerin interaction provides information about structures of type II cohesins and dockerins, as well as their complex state. For cohe-

sins, both type I and type II cohesins share the same overall topology whereas; type II cohesins have additional secondary structure elements. These type II specific elements are thought to contribute to specificity of type II interaction [37]. Correspondingly, type II dockerin displays a considerable similarity with its type I counterpart with several varieties which contributes incisive specificity. As stated below, both helix-1 and helix-3 in type I dockerin can interact with cohesin ligand, alternatively [16]. On the other hand, type II dockerins contact the entire length of cohesin surface with both of its helices. In terms of interaction, the electrostatic surface potentials display variety between type I and type II interactions. The type II interacting interface is less charged than its corresponding type I region, exposing a more hydrophobic nature [33].

2.1.3.2.3 Type III Cohesin-Dockerin Interaction

When the cellulosome assembly in Ruminococcus flavefaciens is identified, the phylogenetic analysis of scaffoldin ScaA and ScaB dockerins expresses a very divergent branch from type I and type II dockerins and classified as type III dockerins [46]. In the course of time, several structural studies demonstrated the distinct construction of type III cohesin-dockerin interaction. Despite its phylogenetic distinction, type III interaction is proved to be Ca^{+2} dependent; similar to type I and type II dockerin-cohesin complexes [15]. On the other hand, in contrast to type I and type II dockerins; type III dockerins lack 22 residue Ca^{+2} binding loop on the second F-hand motif; which is thought to contribute discrepancies in Ca^{+2} binding characteristics and target specificity. Although, recently it is evidenced that Ca^{+2} binding induced similar structural transitions as in type I and type II; the exact structural and biophysical properties of type III cohesin-dockerin interaction is yet to be known [47].

2.1.3.2.4 Dockerin-Cohesin Interaction in Non-cellulosomal Systems

For many years, cohesin and dockerin modules are thought to be elements of cellulosome complex. Thence, it is surprising when these domains are discovered in Archaeoglobus fulgidus, a microorganism that lacks cellulosome [48]. Several other researches prove that non-cellulosomal dockerin-cohesin domains existed in various other bacteria, archaea and in primitive eukaryotes. Interestingly, in about a quarter of the Archaea and 60% of the Bacteria cohesins and dockerins do not co-exist as a pair, one or the other of

the module is missing and the exact role of the modules is these species is not very clearly known [14, 49].

2.1.3.3 Variety in Cellulosomal Systems in Different Bacteria

In 1983, the cellulosome concept is first identified in a gram-positive bacterium, Clostridium thermocellum [50]. To date, cellulosome systems in several other bacteria are revealed, exhibiting diversified cellulosome systems (Figure 2.3). Majority of the bacteria with identified cellulosome systems belong to the genus Clostridium, which are anaerobic and gram-positive [51, 52].

2.1.3.3.1 Clostridium cellulovorans

C. cellulovorans bacterium, possess a simple cellulosome system. Its scaffoldin named CbpA; contains 9 type I cohesin domains and it interacts with several enzymes, mostly glycoside hydrolases [53].

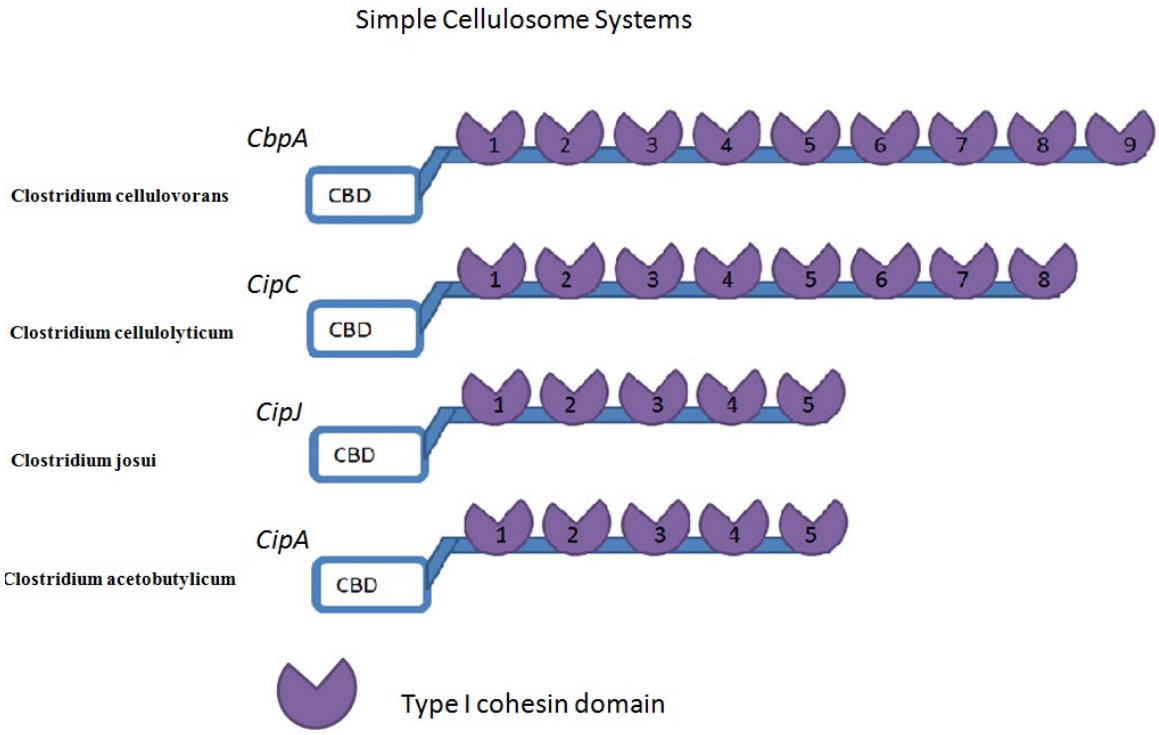

Figure 2.3 Simple cellulosome systems in different bacteria

2.1.3.3.2 Clostridium cellulolyticum

C. cellulolyticum is another anaerobic bacterium that owns a simple cellulosome system. Its scaffoldin is termed as CipC and it has the capacity to interact with up to 8 type I dockerin-borne cellulosomal enzymes [54, 55].

2.1.3.3.3 Clostridium josui

In, C. josui, a simple cellulosome system is organized around a scaffoldin protein named CipA, which bears six consecutive type I cohesin domains [56, 57].

2.1.3.3.4 Clostridium acetobutylicum

C. acetobutylicum, a bacterium with a simple cellulosome system; holds a scaffoldin protein named CipA, which comprises five type I cohesin domains with the ability to bind different cellulosomal catalytic components [51].

2.1.3.3.5 Clostridium thermocellum

C. thermocellum, the first bacterium discovered to have a cellulosome system; features a complex cellulosome structure[50]. The primary scaffoldin called CipA, contains nine type I cohesin domains to recruit type I dockerin-borne enzymes into the cellulosome complex, in addition to its C-terminal type II dockerin domain. Through that type II dockerin domain, CipA interacts with several anchoring scaffoldins that attaches the cellulosome complex to the cell surface [12, 57] . C. thermocellum cellulosome includes three different type II cohesin bearing anchoring scaffoldins; SdbA, Orf2p and OlpB [16]. SdbA, Orf2p and OlpB, carries one, two and seven cohesin domains, respectively [58]. (Figure 2.4)

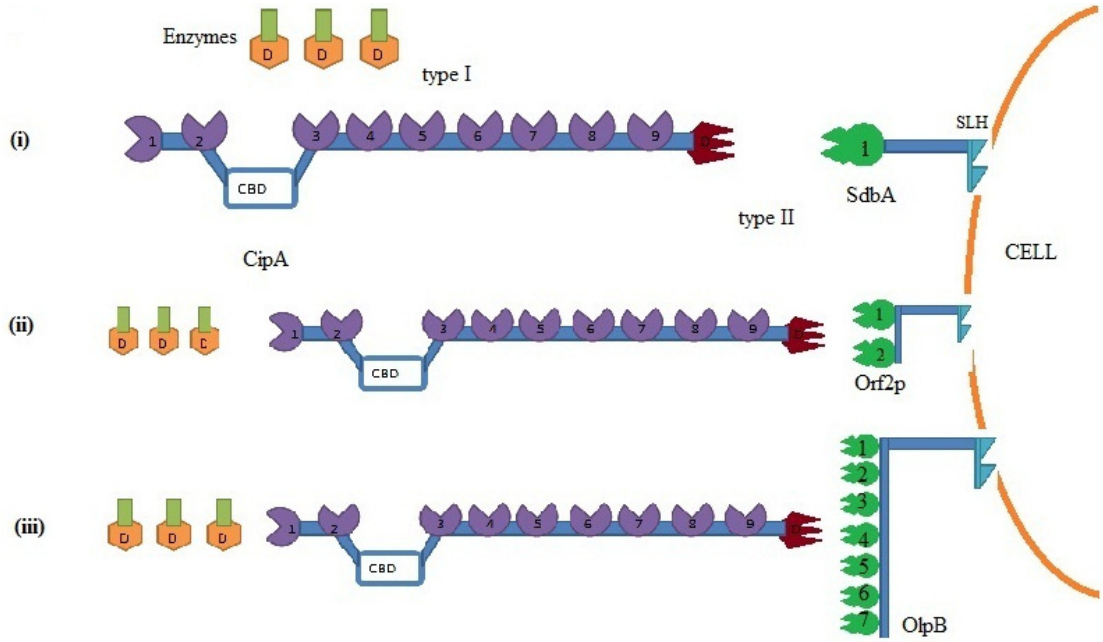

Figure 2.4 Complex cellulosome systems in Clostridium thermocellum.

2.1.3.3.6 Acetivibrio cellulolyticus

A.cellulolyticus is an anaerobic, gram-positive bacterium that displays a complex cellulosome assembly [38]. Its primary scaffoldin is named ScaA and tethers to cell surface via two different mechanisms. Through its C-terminal type II dockerin domain, ScaA can directly bind to ScaD scaffoldin; which is anchored to the cell surface via its SLH module. In addition to the two type II cohesins that interact with ScaA dockerin, ScaD also contains one type I cohesin module that can directly bind enzymes and recruit them to cell surface [59]. Alternatively, through type II cohesin-dockerin interaction, ScaA can bind to the ScaB adaptor protein, which is then attached to type I cohesins of ScaC anchoring scaffoldin [60]. (Figure 2.5, a)

2.1.3.3.7 Bacteroides cellulosolvens

B. cellulosolvens bacterium displays a complex cellulosome system and owns a primary scaffoldin named ScaA; which has 11 type II cohesin subunits to gather catalytic units into the cellulosome complex. Additionally, through its C-terminal type I dockerin sub-

unit, ScaA interacts with ScaB; an anchoring scaffoldin that contains 10 type I cohesin domains. It is a unique example of switched role of cohesin types [40]. (Figure 2.5, b)

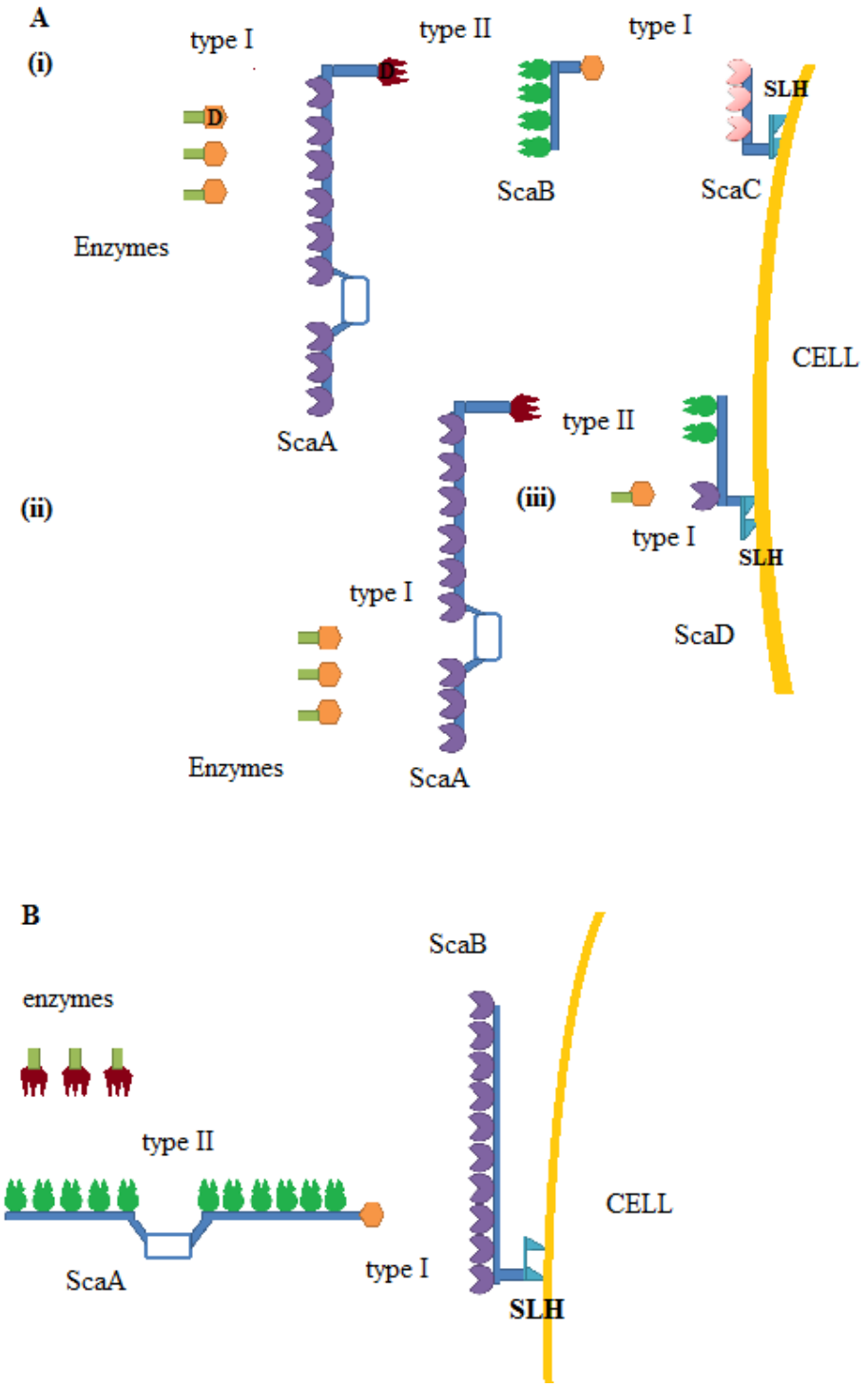

Figure 2.5 Complex cellulosome systems in different bacteria (a) Acetivibrio cellulolyticus (b) Bacteroides cellulosolvens

2.1.3.3.8 Ruminococcus flavefaciens

R. flavefaciens displays two distinct mechanisms that localize dockerin-borne enzymes to the cell surface. In its complex cellulosome structure, a primary scaffoldin named ScaA, interacts with enzymes through its three type I cohesin domain. Alternatively, enzymes interact with only type I cohesin of ScaC and then, ScaC type I dockerin is bound to one of the ScaA cohesins [15]. Subsequently, ScaA-ScaC or ScaA-enzyme complex is localized to ScaB adaptor scaffoldin by the mediation of distinct type III cohesin-dockerin interaction. The enzyme-scaffoldin complex then is tethered to the cell surface via type III interaction of ScaB dockerin and cohesin of ScaE anchoring scaffoldin [47, 61]. In addition to cohesin-dockerin interactions between enzymes and scaffoldins, another type III dockerin containing scaffoldin named cttA is adhered to ScaE. cttA has two CBD domains which coordinates the binding of cellulose, as not other scaffoldins comprise CBDs [62]. (Figure 2.6)

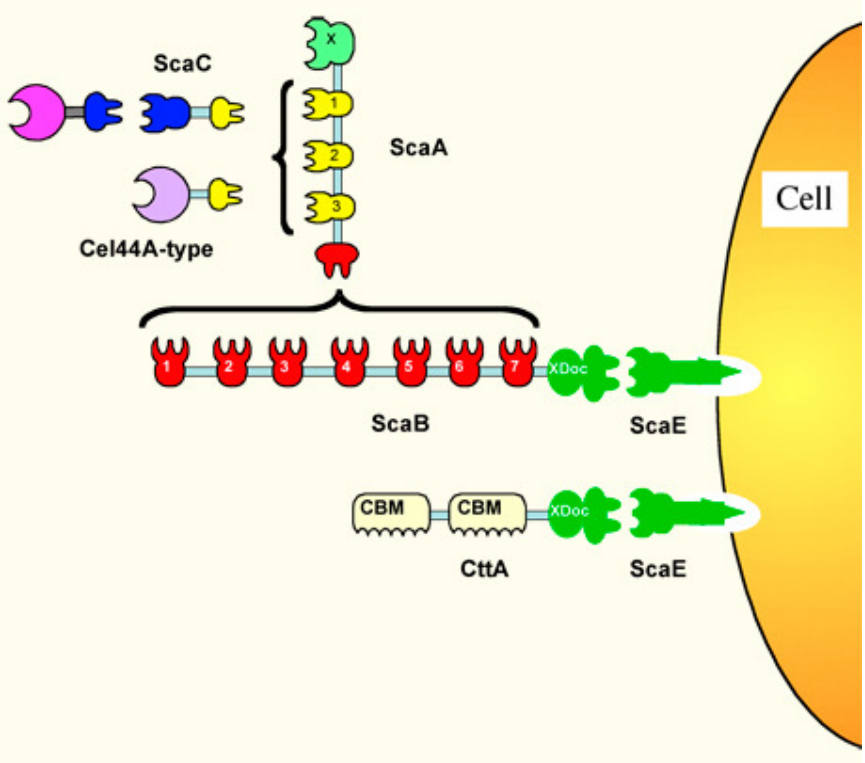

Figure 2.6 Complex cellulosome systems in R. flavefaciens[47]

2.2 Computational Background

2.2.1 Computational Classification Methods

In recent years, a huge number of new cohesins and dockerins are discovered and a large amount of protein sequences is now available in several databases. Structure and functional properties are veiled in mystery for majority of these newly identified proteins [63]. Since the experimental characterization of proteins requires time, effort and is expensive; it is an important task for researchers in bioinformatics to develop computational methods to classify newly identified proteins and predict their function and structure [64, 65].

Hidden Markov Models (HMM), which are extensions of Markov chains; are one of the tools commonly employed in protein classification. In biological context, based on multiple sequence alignments as training set, a HMM calculates similarity scores for new sequences given to the model [66]. In addition to HMM, Support Vector machines (SVM) are another distinguished technique utilized in classification. As an alignment free method, SVM classification tools analyze physicochemical properties of a protein from its sequence. In the presence of sufficient samples from a functional class, SVMs can be trained and classify new proteins against that class, even though the proteins are distantly related [67].

Despite both HMM and SVM methods suggests highly accurate classifications tools, they are unable to determine key-site candidates for interaction. Intended to determine if a given sequence is a member of the training set, HMM techniques are very opaque and it fails to differentiate the key-site candidates [68]. SVMs on the other hand, do not include protein sequence in the classification method, but utilize physicochemical properties derived from dipeptide composition of proteins. In consequence, SVMs do not provide any information on key-site candidates [67]. Additionally, even though SVMs use physicochemical properties, it is impossible to precisely determine which physicochemical properties are significant for interactions and functions [68].

2.2.1.1 Frequently used Protein Classification Methods

In this section, most widely used classification methods, profile HMMs and SVMs are explained in detail.

2.2.1.1.1 Profile Hidden Markov Models

Hidden Markov Models (HMM) are one of the most-preferred protein classification methods. In general, hidden Markov models define probability distributions over a potentially infinite number of sequences [69]. HMMs are extensions of Markov chains. In Markov chains, the choice of the next state is dependent on the current and all state transitions are known, revealing a unique path through the model. However, in hidden Markov models; the state sequence is not observed, it is hidden [70].

HMMs are defined on a finite number of sets (s_1 ,….,s_n), including a begin state and an end state. In order to completely determine an HMM, there are two required sets of probabilities associated with the states:

(1) The transition probability, $T_{i,j}$: For each pair of s_i, s_j states of A, the probability that A will be in the state of s_j at time t+1, given that A is in the state of s_i at time t; where j=i+1,….,n.
(2) The emission probability (output probability) *E(x|j):* For each state s_i, the probability that a particular output symbol is observed in that state. Emission probabilities are properties of only HMMs and not Markov chains. [71]

A 'profile' is a primary structure model based on position specific residue scores and penalties for insertions or deletions. Profile methods use the information in either multiple sequence alignments of structures [72]. The existence of many free parameters in profile methods, such as setting residue scores and penalties, complicates these methods. In order to overcome this kind of problems, hidden Markov models have been introduced to profile methods [73]. Profile hidden Markov models now facilitate several strong tools for protein classification and are employed by several databases [74].

Profile HMMs are probabilistic models that use multiple sequence alignments of a family. Profile HMMs are trained on a representative set of multiple alignments from the family, known as *seed alignments*, to build an HMM profile [66]. For each column in the multiple alignments, match state models the distribution of allowed residues in the column, whereas insert and delete represents insertions of residues between that column and next. Afterwards, to determine if a new sequence is a member of this family or not, its probability to occur by chance is computed using HMM, named E-value. In the cases that E-value is less than a certain threshold, the new sequence is classified as a member

of the family [66, 73]. A schematic representation of profile HMMs is seen in Figure 2.7.

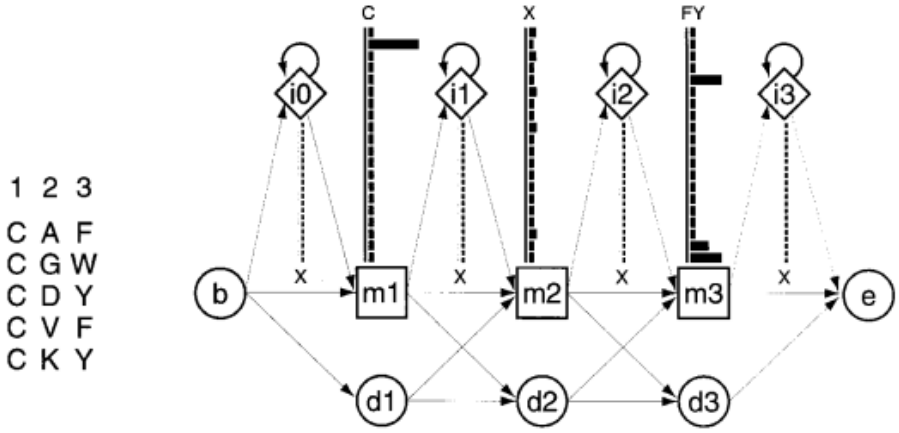

Figure 2.7 A small profile HMM representing the MSA of five sequences (right). The three columns are modeled by three match state (m1-m3), insert state (i0-i3) and delete state (d1-d3). Match and insert states have 20 emission probabilities shown as black bars. Delete states are mute states, with no emission probability. A begin and end state is represented (b,e). Arrows show state transition probabilities [21]

2.2.1.1.2 Support Vector Machines

Support vector machines (SVM) are one of the best discriminative protein classification methods. In brief, SVMs are algorithms that learn how to assign labels to objects [75]. In technical details, SVMs take the input space with nonlinear class boundaries and transforming the input to a new higher dimensional space; they create a linear model to find a plane that separate the positive and negative sets perfectly (Figure 2.8, a). The linear model created by SVMs after transformation is named the maximum margin hyperplane. The maximum margin hyperplane describes a straight line that gives the greatest separation between two, linearly-separable classes. (Figure 2.8, b). The instances closest to the maximum margin hyperplane are then named as support vectors; which define the maximum margin hyperplane for learning. In order to avoid over fitting, in other words too much decision flexibility, usually a few number of support vectors are utilized for hyperplane construction [76, 77].

Unlike homology based methods, SVMs analyze physiochemical properties of a protein generated from its sequence, instead of directly analyzing sequence similarities. Before implementation of SVMs for protein classification, SVMs are used in fold recognition

successfully. Proteins in a specific class generally perform similar functions and thus, share common structural features essential for their function. The structural features directing protein folding are thus anticipated to contribute protein classification [78, 79].

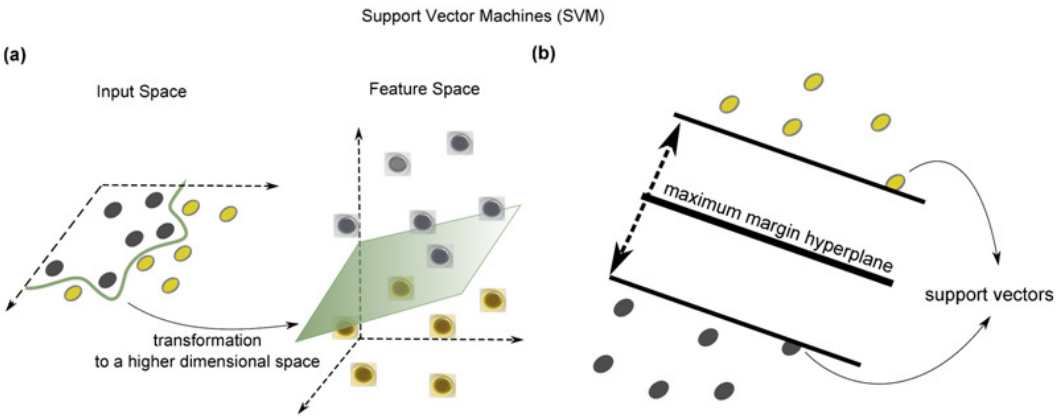

Figure 2.8 (a) The algorithm to find a boundary that maximizes the distance between groups. The input data in two-dimensions cannot be separated by a straight line. The two-dimensional space is transformed into a three dimensional space to separate the data using a hyperplane. (b) The data that are closest to the maximum margin hyperplane are called support vectors. A unique set of support vectors defines the maximum margin hyperplane for the learning problem [83].

The residue properties of proteins might reveal function-related features and construction of an appropriate feature vectors is a key step for successful SVM based protein classification. SVM method utilizes feature vectors constructed from tabulated residue properties, such as amino acid composition, charge, hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure, solvent accessibility and surface tension [80, 81]. Independent of sequence similarity, this approach is capable of classifying distantly related proteins with low sequence homology as well as the highly similar related proteins [79].

2.2.2 Biological Aspects of Protein Classification Problem

In order to understand the biological processes, knowledge about the functions of proteins is of great importance [67]. The recent revolutions in high-throughput technologies facilitates to procure information about the structure and function of biological

molecules [82]. Benefiting from the advantages of high-throughput technologies, several genome projects revealed a vast amount of sequences information for a large number of organisms. The rapid accumulation of sequence information lead the scientist to develop new methods for protein function prediction from sequence because, the functions remain unknown for the majority of the newly identified sequences [64, 65].

Experimental characterization of protein functions is a valuable source to understand how these proteins function in a living organism. On the other hand, experimental methods may be high-cost and time-consuming. Hence, several computational methods are developed for reliable and large-scale protein function annotation, cooperated with experimentally verified information [64, 65]. In order to obtain clues about the function and interactions of proteins, a meaningful classification linked to existing experimental knowledge is necessary.

In brief, classification methods identify the similarities (homologies) between protein sequences and group them into particular classes. In technical terms, classification basically requires the collection of certain components. The first component required for classification is the elements to be classified such as protein function and structure. Subsequently, certain characteristics of these elements are defined to be used in classification and based on these characteristics, a similarity or distance metric is derived. Another component in classification is the algorithms to generate metrics and build clustering and classification. Finally, interpretation of relationships between clusters; which is linked to the performance evaluation of the entire procedure terminates the classification process [82].

Subfamily identification, division of dataset into subclasses, offers several advantages for classification methods. Existence of a structurally characterized member in a subfamily enables to render an opinion about the structure and function for other members of the subfamily. Additionally, identification of known subfamilies facilitates the usage of support vector machines (SVM) and sequence based classification methods to classify indeterminate sequences into existing subtypes [65]. In order to perform sequence based subfamily classification, several statistical models that employ the information in multiple sequence alignments have been developed, such as profiles and hidden Markov models (HMMs) [83]. In addition, various SVM based discriminative classifiers that appoint unlabeled proteins into predefined subfamilies are designed [84]. The basic

principles of these approaches will be discussed precisely in the following parts of the thesis.

2.2.2.1 Homology Detection Approaches

In the classification problem, the methods to detect similarities between sequences can be divided into three basic groups:

Pairwise Sequence Comparison Algorithms: The most popular sequence comparison methods in this group are BLAST and Smith-Waterman (SW). The SW algorithm utilizes dynamic programming to produce an optimal local alignment between two sequences [85], whereas BLAST calculates a heuristic alignment score to approximate SW [86].

Generative Models: These models are trained on datasets and represent positive features of a protein family. Based on the extracted features, close homologs are added into a positive group and classified into that family. Profile HMM method is one of the most widely-known generative models [73, 87].

Discriminative Classifiers: In this method, classifiers such as SVMs are trained on both positive and negative data to distinguish between classes [87, 88].

Based on different homology detection approaches, scientist develops several protein classification methods.

2.2.3 Reduced Amino Acid Alphabets

Reducing the 20-letter amino acid alphabet into a smaller number by grouping similar amino acids together is an effective approach utilized with protein classification methods. A variety of such amino acid groupings called reduced amino acid alphabets are defined and tested for classification efficiency [68]. The utilization of RAAAs can also pinpoint key site candidates conserved in terms of amino acid property.

As stated many before; functional, structural and many other biologically relevant information for the newly identified sequences can be inferred from the evolutionary related sequences by computational methods. For most of these methods, the sequence

alignment is a standard method. Even though the accuracy of the alignments is of significant importance, the substitution matrices used for alignment have considerable impact upon the reliability of the alignments [89, 90].

Most of the popular substitution matrices such as PAM, BLOSSUM and GONNET are build based on sequence alignments and unfortunately, the accuracy of the alignments, and therefore the substitution matrices, become less reliable for the distantly related, low-sequence similarity sequences [91]. In an effort to dispose the problems resulting from sequence similarity issues, several solutions have been proposed by scientists. Amino acid grouping based on similarity is one of the major adopted solutions and the amino acid alphabets produced by these groupings are named '*reduced amino acid alphabets'* (RAAA).

The proper groupings of amino acids reveal the similarities which are invisible in the full 20-letter alphabet and ensure statistical significance in applications of protein bioinformatics, such as structure prediction, homology detection and functional classification [92]. However, the compression of amino acids also causes the loss of certain amount of information. Therefore, the balance between maximal conservation of information and statistical significance is of cardinal importance [93].

A variety of amino acid grouping schemes is suggested utilizing different similarity measures. Groupings based on structural alignments and physiochemical properties are the most widely-used ones. Structural features of proteins from the same functional classes are more conserved than their sequences and the structural alignments are reliable even for proteins in distant evolutionary relationships. Depending on the distribution of amino acids in structural units, several structure-based similarity matrices have been developed. Based on those similarities, different amino acid groupings have been proposed such as GMBR, HSDM and SDM [91, 93].

Amino acid groupings based on the physiochemical properties is another well-known approach adopted for RAAAs. During evolution, mutations that do not change physical and chemical properties of the amino acids are accepted, even in the conserved sites, since the function of the molecule is not disrupted by these mutations [94]. Numerous methods attempted to group amino acids based on their different physiochemical properties and various RAAAs are defined [95-97].

2.2.4 Correlated Mutations

The positive Darwinian selection is a mode of natural selection that favors some alleles, on the contrary of negative selection that removes the lethal or disfavored alleles. Despite the rare occurrence of positive selection process; the fragments responsible for biologic activities such as reactive sites and interaction sites are more prone to positive selection [98] .

The current model of positive selection assumes that positive mutations occur in an interconnected manner. The changes occurring in the neighborhood are related to the positive mutations and generally, these interconnected changes reflect protein interactions, biological activity and structurally significant units of the molecule. Therefore, the fixed mutations related with each other should occur concurrently [99]. The simultaneous occurrence of several mutations is known as *correlated mutations*. The relationship between correlated mutations and the role of the involved sequences in protein-protein contacts are demonstrated by several reports [100, 101].

The correlated mutations phenomenon is not constrained with intra-protein residues and can be expanded to inter-protein interactions. On the interacting surfaces of proteins, the amino acid substitutions are more limited because of the functional and structural constraints. However, once a significant residue for interaction is changed, the effect of the functional constraints on the interaction surface can be counterbalanced by an additional mutation on a complementary residue. The coevolution of two proteins can lead high specificity and affinity [102].

Although the fundamental idea behind the concept of correlated mutations has a straightforward nature, establishing and quantifying is a challenging task. The methods proposed for correlated mutation studies are still not very diversified [102]. The most widely-known approaches for correlated mutation analysis include McBASC [103] , OMES [104], MI [105], Quartets [106], ELSC [107] two-state maximum likelihood methods [108].

Chapter 3

METHODOLOGY

3.1 Introduction

Conserved residues in protein sequences are often found to be consequential for protein-protein and protein-ligand interactions. In some cases, however, instead of a specific residue, some physiochemical properties of amino acids are conserved. Type I, type II and type III dockerin-cohesin interactions differ in terms of their interaction structure and conserved residues in one subclass can designate the mandatory residues for proper function and structure. In our approach, we benefit from conserved residues to classify dockerins and cohesins. In order to utilize the information present in conserved properties, several reduced amino acid alphabets are introduced. In this study, every step is conducted on each alphabet. At the initial stage, sequences in each subclass are aligned separately. In order to pinpoint the residues that serve as motifs, residues conserved only in one subclass but not in others are detected. All of the detected residues are ranked with a scoring function which measures the specificity for their subclass. Then, the residues with high distinguishing capacity are selected as motifs and used for classification.

Another aspect of our study is to target candidate key residues for different types of dockerin-cohesin interactions. Each motif utilized for classification, are also treated as candidates for key interaction sites. It is reported many times that residues directly contact in protein-protein interactions overlap with correlated mutation studies. In order to affirm the candidate key site residues, each subclass are surveyed for correlated muta-

tions. Figure 3.1 depicts a schematic illustration of our method and each step is explained in detail, in the following sections.

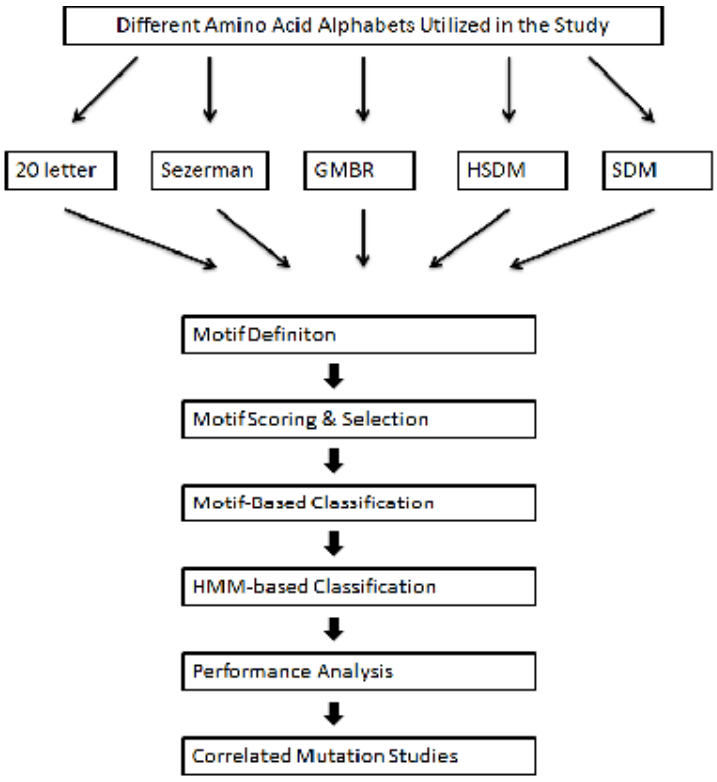

Figure 3.1 A schematic representation of the methodology.

3.2 Data Collection

All our experiments are performed on a set of proteins. Training and test data sets needs to be defined before implementing any algorithm. For both cohesin and dockerin sequences, training and test sets are prepared. All methods used in this work are trained on the training set at first, and their classification performance is tested against test sets. In these test sets, each sequence serve as a positive test sequence for its own class and negative test sequence for the other classes.

In addition, a set of sequences with unknown subclass are classified using the proposed method. This set is independent of train and test sets and is not utilized for motif selection.

3.2.1 Data Sources

Both dockerin and cohesin sequences used in this study are retrieved from UniProt KnowledgeBase (UniProtKB) database. UniProtKB is a database under UniProt, which is an extensive protein sequence and annotation data resource. UniProtKB provides a collection of sequences and functional information on proteins. UniProtKB is composed of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/TrEMBL entries are derived from computationally generated, hypothetical translation of coding sequences (CDS), whereas; UniProtKB/Swiss-Prot brings computed features and experimental results together providing high-quality, non-redundant protein sequences [109].

All available dockerin and cohesin sequences in UniProtKB are extracted in order to utilize in this study. After extraction, the sequences are divided into three, based on their types. At that point, we obtain three different datasets for both cohesins and dockerin sequences, named as type I, type II and type III.

3.2.2 Training and Test Data

Following the sequence extraction, training and test sets for each subclass (type I, type II, type III) are defined. Type I training and test sets for both dockerins and cohesins are defined by randomly dividing the type I datasets in 1:1 ratio. Thereby, cohesin type I train set contains 36 sequences and dockerin type I train set contains 68 sequences. For cohesins, type II and Type III train and test sets are defined by the same way, and they include 22 and 5 sequences, respectively. For dockerins, since there are limited number of type II and type III sequences, train sets of these subtypes include 3 sequences. Since we have only 3 type II sequences, there is no type II test set and type III test set includes one sequence. Each training set is used as positive training set for its own class of proteins, whilst the other training sets serve as negative training sets for that class.

3.2.3 Data with Reduced Amino Acid alphabets

Despite the availability of large number of possible amino acid sequences, the number of folds that proteins can hold is comparatively low. In some cases, sequences that display almost zero identity can adopt considerably similar structures. This degeneracy has

lead to development of reduced amino-acid alphabets; new sequence descriptions that have the capability to reveal structural similarities of dissimilar sequences [91].

In general, reduced amino-acid alphabets can be defined as grouping of amino acids based on measures of their relative similarity. Peterson et al. compared the sensitivity and selectivity of several reduced amino-acid alphabets[92] . The top performing three RAAAs from this comparative work are selected to be utilized: GMBR, HSDM and SDM alphabets. In addition, Sezerman's amino-acid grouping is also employed.

GMBR alphabet is defined by Solis et. al, based on local structure coding behavior similarities. Without regard to any physiochemical variables or multiple sequence alignment of homologous structures, GMBR alphabet is set solely according to the similarities of amino-acid distributions in local structures [93]. Based on the relative propensities of amino-acids for structures, amino-acids are grouped into a four-letter alphabet, detailed in Table 3.1.

Table 3.1 Amino acid groupings utilized in this study

Alphabet	Amino Acid Groups															
GMBR	ATSQNDEKR								HYFLIVMCW						G	P
	A								*Y*						*G*	*P*
Sezerman	A	TS		QN		DE		KRH	YF		LIVM		C	W	G	P
	A	*T*		*Q*		*D*		*K*	*Y*		*L*		*C*	*W*	*G*	*P*
SDM	A	TSQ			N	D	EKR		YF		LIVM		C	W	G	P
	A	*T*			*N*	*D*	*E*		*Y*		*L*		*C*	*W*	*G*	*P*
HSDM	A	T	S	Q	N	D	EK	R	Y	F	LIV	M	C	W	G	P
	A	*T*	*S*	*Q*	*N*	*D*	*E*	*R*	*Y*	*F*	*L*	*M*	*C*	*W*	*G*	*P*

HSDM and SDM reduced alphabets are also defined by homologous structure alignments of proteins with low-sequence similarity; yet less amino-acid substitutions are introduced as compared to GMBR. The main difference between HSDM and SDM alphabets is their protein datasets that have been used for structural alignment. SDM alphabet is calculated based on the structural alignment of all protein sequences in their dataset. HSDM alphabet, on the other hand, is derived from the dataset after proteins of unclear evolutionary relationships are excluded [91]. Amino-acid groupings of HSDM and SDM alphabets are summarized in Table 3.1.

In Sezerman alphabet, amino-acids are grouped according to their physical and chemical properties, unlike GMBR, HSDM and SDM alphabets, in which groupings are based

on structural alignment and propensities of amino-acids for different structural units [94]. The amino-acid grouping in Sezerman alphabet is detailed in Table 3.1.

Following the decision of test and training datasets, the sequences are translated into each of the RAAAs explained above. The rest of the studies carried out in this thesis, are conducted on each of these alphabets, separately.

3.3 Protein Classification

The purpose of this thesis is mainly to classify dockerin and cohesin sequences into their subtypes and determine key site candidates for their interaction. To this end, we utilize a classification technique combined with RAAAs that targets key interaction site candidates. In this method, the residues that are conserved highly in one subclass but show slight conservation in others are identified as motifs and given scores according to their specificity among classes. These motif scores are then used to calculate a total score for each subclass to decide on a classification rule. These rules are applied to test sequences to classify them into subclasses.

3.3.1 Motif Definition

During evolution, the function of a protein is predominantly conferred by small parts that form the critical regions, like active sites and binding sites [110]. Those key residues are shown to be correlated with the most conserved amino acids in proteins. Proteins have undergone several changes in their sequences throughout the time, however; maintenance of structural integrity and function entails conservation of the key residues, or at least their property. These highly conserved fragments with significant biological importance are called as 'motifs' [111].

Motifs have been wildly utilized for structure prediction and protein classification studies [65]. Motif based classification approaches have been shown to discriminate even between highly similar sequences [112]. In addition, single residue conservation in proteins is often found to be significant for protein-protein interactions. Binding surfaces of proteins are subjected to strong selective pressure and therefore, conserved residues have been candidates for binding sites [113]. Despite their high classification efficiency, in motif-based classification methods, single residue conservations among classes may

be overlooked. In this thesis, as well as the correct classification of dockerin and cohesin sequences, the key site candidates for the interaction are aimed to be determined. Therefore, conserved residues are increased in importance. In the frame of classification, conserved single residues can also proffer suggestions simply because the cohesin-dockerin subclasses are distinguished by diversities in interactions. On these grounds, we conduct our classification method on conserved single residues. As a matter of convenience, the conserved single residues will be referred as motifs throughout the thesis.

In order to determine motifs, sequences from each subclass of dockerins and cohesins are aligned separately. The consensus sequences of 70% conservation are then utilized for motif definition. The residues that present in the target subclass but not in others are defined as motifs for that specific subclass, for all amino acid alphabets.

3.3.2 Motif Selection and Scoring

In the method above, the residues conserved in only one subclass are defined as motifs. However, it is probable that a non-conserved residue is present in some sequences of a subclass, with conservation less than 70%. Hence, each motif owns a different specificity and low specificity motifs should be removed prior to classification. An ideal motif would be one that is observed in all sequences of its subclass and never occurred in sequences of others. In order to determine closeness of motifs to the ideal case, a scoring function that gives high scores to the motifs frequent in the target subclass and rare in others is applied, named Motif Specificity Score (MSS).

Motif specificity score is composed of two parameters: Presence in Subclass and Subclass Specificity. Presence in Subclass is a measure of motif's occurrence in target subclass. Presence of motif *i* in subclass *j,* $PS_{i,j}$, is defined as:

$$PS_{i,j} = \frac{n_{i,j}}{S_j} \tag{3.1}$$

Where $n_{i,j}$ is the number of sequences containing motif *i* in subclass *j* and S_j is the total number or sequences in subclass *j*.

The second parameter of motif specificity score, subclass specificity SS_i, is in inverse proportion with the number of sequences containing motif *i* from other subclasses. SS_i is defines as follows:

$$SS_{i,j} = e^{-(\frac{n_{i,k}}{S_k})} \tag{3.2}$$

Where $n_{i,k}$ denotes the number of motif *i* in the subclasses other than *j* and S_k is the total number of sequences in those subclasses.

Following the calculation of PS and SS, MSS of motif *i* in subclass *j* is calculated as follows:

$$MSS_{i,j} = PS_{i,j}\mathbf{x}SS_{i,j}\mathbf{x}100 \tag{3.3}$$

The next step following the calculation of motif specificity scores is determination of a certain threshold value to filter motifs with high distinguishing capacity and utilize them for classification. The distribution of subclass specificity and motif specificity scores displays varieties between cohesin and dockerins. Hence, different threshold values are set. Selected motifs and their MSSs for each amino acid alphabet is provided in Appendix A.

3.3.2.1 Cohesin Sequences

In the motif definition step, a threshold value for presence in subclass (PS) score is defined, 70. A high percentage of motifs identified from cohesin sequences show perfect subclass specificity, with $SS_{i,j}$ value 1. Due to the presence of highest specificity level motifs, instead of defining a MSS threshold, different thresholds for PS and SS are set. The SS threshold is set to 1, indicating that motifs occurred only in the target subclass are used for classification of cohesin subclasses.

3.3.2.2 Dockerin Sequences

Unlike the maximum level subclass specificity of cohesin sequences, the number of motifs with maximum subclass specificity defined for dockerins is not sufficient for an accurate classification. In the case that a small number of motifs are set for prediction, on the other hand, the risk of misclassification increases. Therefore, the threshold value

for subclass specificity is set to a lower value, 0.9, to obtain higher number of motifs. The PS score threshold is again set as 70.

3.3.3 Motif Based Classification

Classification based on motifs is then implemented on the test set. Each of the sequences in the test set is first aligned to the profile alignment of the target subclass, separately. Subsequently, the sequences are inspected to expose whether the motifs of the target subclass are present or not. Based on the specificity scores of motifs occurred in the sequences, a Classification Score (CS) is defined for each test sequence, as follows:

$$CS_s = \frac{\sum_{i \epsilon N} MSS_{i,s}}{\sum_{i \epsilon F} MSS_i} \tag{3.4}$$

In the numerator, $MSS_{i,\ s}$ indicates the motif specificity score of motif *i* in sequence *s* and *N* is the set of target subclass motifs present in sequence *s*. In the denominator, MSS_i exhibits the motif specificity score of motif *i* where *F* is the set of all motifs for the target subclass.

The division with the motif specificity score of all motifs for the target subclass functions as a normalization step. Without normalization, the subclass with more number of total motifs can dominate the score, even if it is not the correct subclass. In order to avoid that kind of a misclassification, we add the division to our CS calculation.

For each subclass, the CS values of test sequences are calculated and each test sequence is assigned to the subclass with the highest CS value.

3.4 Classification with profile HMM

In order to compare the performance of our method with state-of-the-art classification methods, a profile HMM classification is conducted on our test set. For each subtype, the process is repeated in 20 letter alphabet, for both cohesin and dockerin.

Sequence Alignment and Modeling (SAM) system software version 3.4 is utilized in this study. SAM is an implementation of profile HMM method for protein classification. Unaligned positive train set sequences in FASTA format is provided as input. In order to obtain better probability distributions, Dirichlet mixture priors are used. Dirichlet mixture priors are introduced to profile HMMs by Sjolander et al. [114]. In this method, multiple sequence alignment information from databases are condensed into a mixture of Dirichlet densities over amino acid distributions and these densities are combined with the observed amino acids to obtain more effective estimates of the expected amino acid distributions. The following commands and options are used for classification:

> buildmodel train_model -train trainset.fas -prior_library uprior.9comp -randseed 0

Here, trainset.fas is the input file and uprior.9comp is the Dirichlet mixtures library. The model built by SAM is named as train_model and it is saved in train_model.mod file. The command randseed is for initial model length selection and setting it to 0 makes the program run reproducible.

Following the construction of model, each test sequence is compared to each of the models, using the commands as follows:

>hmmscore outfile -i train_model.mod -db testset.fas -sw 2 -calibrate 1

The unaligned test sequences are available in FASTA format in testset.fas file. The outfile parameter sets the name of the output file (outfile.dist). For each of the test sequences, E-values are calculated based on the model, given as train_model.mod and e-values are contained in outfile.dist file. –sw parameter sets the type of the alignment and setting it to 2, we performed full local alignments of sequences to the model. –calibrate parameter set to 1 to obtain better calibration of e-values. In order to complete classification, a certain threshold for e-values is set and the sequences with the e-values lower than or equal to threshold is classified as a member or target subtype (positives). The sequences with e-values higher than the threshold are labeled as negatives of that target subtype. The threshold for e-values is decided using the minimum error point approach, described in section 3.6.3.

3.5 Performance Analysis

Classification performance of the methods is analyzed using various statistics. Firstly, cross-validation studies are conducted. Gini Index is introduced to observe how significant the discrimination between type I, type II and type III scores is. The confidence level of scoring is then calculated. Minimum error point is utilized to define a threshold value for HMM classification and lastly; confusion matrix, accuracy rates, sensitivity and specificity values of classifications are analyzed to compare our method with profile HMM.

3.5.1 2-fold Cross-Validation

2-fold cross-validation analysis is performed for all experiments in this study. For cross-validation; using the same dataset, different train and test sets are defined randomly dividing the dataset into two, as described before. Differently, for dockerin type II set that has no test sequence, one of the sequences in train set is left outside from motif definition and treated as a test set in cross-validation studies. Therefore, the algorithm is trained and tested on different sets. This process is repeated twice. For rest of the thesis, the first study conducted is named as the study1, whereas the cross-validation studies are named as study2 and study3.

3.5.2 Gini Index

Gini index is a statistical approach that have been widely used in economy to evaluate the distributional properties of income and wealth [115]. In biology, Gini index has been used for evaluation of computational classification methods to rate the weight of the features used for classification. In summary, the Gini index measures the impurity/inequality of the samples that are split from a common parent node [115].

A smaller Gini index implies higher purity, in other words, best separation between the nodes. In the case of equal distribution at a node, the Gini index gains the maximum value at $1\text{-}1/n_c$,implying least differential information. n_c is the number of classes at the mode. On the other hand, Gini index gets its minimum value, 0, where all information comes from only one node, implying the most differential information [116]. The Gini index is computed as follows:

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2 \quad (3.5)$$

Where *p(j|t)* is the relative frequency of class *j* at node *t*. Calculation of Gini index is demonstrated in Table 3.2.

Table 3.2 Calculation of Gini Index

Classes at the Node	Values assigned to classes	Calculation
C1	5	p(C1)=5/5=1 p(C2)=0/5=0
C2	0	Gini=1-p(C1)^2-p(C2)^2=1-1-0=0

Classes at the Node	Values assigned to classes	Calculation
C1	7	p(C1)=7/10=0.7 p(C2)=3/10=0.3
C2	3	Gini=1-p(C1)^2-p(C2)^2=1-0.49-0.09=0.42

In our classification step, the type with the highest score is assigned as the type label of the target sequences. In an ideal case, the score of the unassigned types should be zero, since the sequence should have no motif residues other than its own type motifs. However, in some cases, the scores of the unassigned types for the target sequence are very close to the score of the assigned type. So, in order to determine how close they are to the ideal case, the Gini index for each test sequence is calculated after the classification. Scores for each type are handled as the classes at a node. In this way, Gini index established how differential the type scores are.

3.5.3 Confidence Intervals

In classification problem, determining how close an estimate to the parameter being estimated is a crucial question. In this thesis, in order to address this question, confidence intervals are introduced to our method.

In statistics, a confidence interval for a population parameter is a range of values defined with a certain confidence that the interval contains the unknown population parameter, the estimated range been calculated from the sample population. The confidence in this concept means the degree of certainty that the unknown parameter belongs to that population [117].

In our classification method, in order to statistically infer how significant a classification score can predict the subclass of the sequence, the confidence intervals of classification scores with 99% confidence level are calculated. Following, the ratio of the test sequences with classification scores higher than the lower end point of the confidence interval are calculated. In order to determine the confidence intervals, the classification scores for training sequences are calculated. Subsequently, the confidence intervals are calculated based on the training set classification scores as the sample population, as follows:

$$[\hat{\mu} - (T)(\frac{\hat{\sigma}}{\sqrt{n}}), \hat{\mu} + (T)(\frac{\hat{\sigma}}{\sqrt{n}})] \tag{3.6}$$

$\hat{\mu}$ is the sample mean, $\hat{\sigma}$ is the sample standard deviation, *n* is the sample size and lastly, *T* is a table value that depends on the confidence level, obtained from T-2 table, in with degree of freedom n-1. As we used 99% confidence level, if the scores smaller than lower end point of the interval are not classified as a member of the target subclass, the probability of erroneously rejecting a sequence of the target subclass is no greater than 0.005.

3.5.4 Minimum Error Point

Minimum error point (MEP) can be defined as the score threshold at what point a classifier makes the minimum mistakes, of both false positives and false negatives [84]. Each classifier outputs scores for their prediction. The test sequences are ranked according to their prediction scores and prediction errors (FP+FN) are calculated as each score is treated as the threshold. The threshold level with the minimum number of errors is the minimum error point (MEP). MEP denotes the best case accuracy of a classifier.

The minimum error point is calculated for profile HMM classification, to determine the threshold value for classification.

3.5.5 Confusion Matrix, Accuracy Rates, Sensitivity and Specificity Calculations

Simply, confusion matrix is a 2x2 table, showing the number of accurate classification and errors, predicted by a classifier. As shown in Table 3.3, a confusion matrix is composed of 4 elements:

-*True Positives (TP):* Number of sequences that are truly predicted to belong to the target subclass.

-*False Positives (FP):* Number of sequences that are falsely predicted to belong to the target subclass.

-*True Negatives (TN):* Number of sequences that are truly predicted not to belong to the target subclass.

-*False Negatives (FN):* Number of sequences that are falsely predicted not to belong to the target subclass.

Following the definition of confusion matrix elements, the accuracy rate is given by:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

Table 3.3 A confusion matrix and its elements: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).

	Predicted Label		
Actual Label		+	-
	+	True Positives (TP)	False Negatives (FN)
	-	False Positives (FP)	True Negative (TN)

In addition to the accuracy rates, sensitivity and specificity levels of classifications are calculated from confusion matrix. Sensitivity and specificity levels are calculated as follows:

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.8}$$

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{3.9}$$

3.6 Correlated Mutations

Correlated mutation studies are conducted using Coevolution Analysis using Protein Sequences (CAPS) software version 1.0. CAPS software measures the Coevolution between amino acids from the same protein (intra-molecular Coevolution) of from two functionally interacting, distinct proteins (inter-molecular Coevolution) [118].

In this study, the inter-molecular coevolution between dockerin and cohesin domains is analyzed. The Caps.ctl file available in CAPS software includes the control parameters for analysis. The Co-evolution analysis parameter in Caps.ctl file is set to 1 to perform an inter-protein analysis. The Input file1 parameter is left as Groel.aln and Input file 2 parameter is changed as Groel2.aln. Type of data 1 and Type of data 2 parameters is set to 0 to work on the amino acid alignments. The other parameters are left as default. Aligned cohesin and dockerin sequences are given to the software in FASTA format as input. The input files are named as Groel.aln and Groel2.aln, respectively. After the changes in caps.ctl file, the following command line is used to conduct the coevolution analysis:

> perl caps.pl

The output file created by the software is named as Groel.out. This study is conducted on the initial train sets of type I, type II and type III subclasses, separately.

Chapter 4

RESULTS AND DISCUSSION

As described in Chapter 3, our classification method is trained on different type I, type II and type III training sets, for cohesins and dockerins. For each train set, four different reduced amino acid alphabets are introduced to analyze effects of groupings in classification. Test sets containing sequence from each subclass are used to test the performance of the proposed method. Subsequently, correlated mutation studies are conducted on training sets. In this chapter, the results of the tests identifying dockerin and cohesin subclasses are discussed firstly. Following, the predictions for the sequences with unknown subclasses are provided. In the second section, the correlated mutation studies and their results are analyzed.

4.1 Identification of Dockerin-Cohesin Subclasses

To examine the performance of proposed classification method on different reduced amino acid alphabets, test sets in each alphabet containing sequences from each subclass are used. In order to analyze the performance, 2-fold cross-validation studies are conducted and cases are named as study1, study2 and study3.

4.1.1 Subclass Identification for Dockerin

The classification method utilized in this study is applied to both cohesin and dockerin sequences, separately. In this section, classification and performance results for dockerin sequence are discussed.

4.1.1.1 Confusion Matrix and Accuracy Rates

The confusion matrix of dockerin classification for each RAAA and each study is summarized in Table 4.1. Type I sequences are generally classified accurately, however in some cases, they are misclassified as type III and in one case, a considerable number of type I sequences is classified as type II. Likely, type II and type III sequences are mostly misclassified to each other and type I. However, the proportion of the misclassified sequences does not show correlation between different RAAAs or different studies.

Table 4.1 The confusion matrix of dockerin classification. In each section, rows represent different RAAAs and columns represent the cases; study 1, study 2 and study 3, respectively.

Confusion Matrix	**Type I Actual**	**Type II Actual**	**Type III Actual**	**Alphabets**
Type I Predicted	65 64 67 66 50 67 67 56 68 65 61 67 64 61 64	0 0 0 2 0 0 0 0 0 0 0 0 0 0 1	0 0 1 0 0 1 0 1 0 0 1 0 0 0 0	20 letter GMBR HSDM SDM Sezerman
Type II Predicted	0 0 0 0 15 0 0 0 0 0 0 0 0 0 0	0 0 1 0 1 1 0 0 1 0 0 1 0 0 0	0 1 0 0 0 0 1 0 0 0 0 0 0 1 0	20 letter GMBR HSDM SDM Sezerman
Type III Predicted	3 3 1 1 3 1 0 12 0 3 7 1 4 7 1	0 1 0 0 0 0 0 1 0 0 1 0 0 1 0	1 0 0 1 0 0 0 0 1 1 0 1 1 0 1	20 letter GMBR HSDM SDM Sezerman

The accuracy rates for the proposed classification method on different studies of five different amino acid alphabets are listed in Table 4.2. It is noteworthy that the average accuracy levels of the method for all amino acid alphabets are higher than 90%. In order to examine the effect of different training and test sets on the classification method, cross-validation tests are conducted. Except the comparably low accuracy rates in study2, the individual accuracy rates are even higher than average rates.

Table 4.2 The accuracy rates and Gini index values of dockerin classification for different amino acid alphabets and for cross-validation studies on different datasets.

Proposed Method	Accuracy Rate (%) *Gini Index*				
	20 letter	GMBR	HSDM	SDM	Sezerman
Study 1	96% *0.612*	96% *0.553*	97% *0.589*	96% *0.574*	94% *0.568*
Study 2	93% *0.587*	76% *0.651*	80% *0.568*	87% *0.579*	89% *0.631*
Study 3	97% *0.634*	97% *0.631*	100% *0.645*	98% *0.644*	89% *0.643*
Average	*95%* *0.611*	*90%* *0.611*	*92%* *0.601*	*94%* *0.599*	*91%* *0.614*

Among the five different amino acids alphabets utilized in this study, the average accuracy levels do not display drastic changes, as summarized in Table 4.2. In average, 20-letter alphabet gives the higher accuracy level, 95% and with 94%, HSDM alphabet follows. However, if we disregard the low accuracy rates in the first cross-validation study, HSDM alphabet gives the highest accuracy rates, 97% and 100%. Following HSDM alphabet; 20-letter, GMBR and SDM alphabets give nearly the same accuracy rates, 96% and 97%, respectively. However, the accuracy rates do not display correlation between different studies or RAAA, as in the confusion matrix values.

Note that due to limited number of available type II and type III dockerin sequences, training sets of these subclasses are in small scale compared to type I train sets. Thence, it is discussible whether the motifs defined by these train sets reflect the true conservation for these subclasses. However, despite the uncertainty of type II and type III motifs, the method gives remarkably high overall accuracy levels, especially for type I prediction. The specificity and sensitivity values based on confusion matrix are calculated and presented in Table 4.3.

As seen in Table 4.3, Type I prediction for all amino acid alphabets gives high sensitivity. However, type I specificity values display drastic changes between different alphabets and different studies. Since we have only one type II and type III test sequences, the sensitivity values may not be very reliable and expectedly, for both of the subclasses

the average sensitivity rates are low. In addition, the sensitivity values are demonstrating significant differences between RAAAs and studies. For specificity however, type II and type III sequences give comparably high values and these values are relatively correlated between amino acid alphabets and studies.

Table 4.3 Dockerin sensitivity and specificity values calculated from confusion matrix for type I, type II and type III prediction on five different amino acid alphabets. Different colors represent different amino acid alphabets; 20-letter, GMBR, HSDM, SDM and Sezerman, respectively.

	Type I		Type II		Type III	
	Sensitivity (%)	**Specificity (%)**	**Sensitivity, (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
Study 1	**96, 96, 99, 96, 94**	**100,100,100, 100,100**	**X**	**X**	**100,100, 0, 100,100**	**96, 99, 100, 96,94**
Study 2	**96, 74, 82, 90,90**	**33, 100, 50 50,100**	**0, 0 , 100, 0, 0**	**100,77,100, 100,99**	**0, 0, 0 0,0**	**94, 96, 81, 88,88**
Study 3	**99, 99, 100, 99,96**	**50, 50, 100 100, 50**	**100,100,100, 100,0**	**100,100,100, 100, 97**	**0, 0, 100 100,100**	**99, 99, 100, 99,96**
Average	**97, 90, 94, 95, 93**	**61, 83, 83, 83, 83**	**50, 50, 100, 50, 0**	**100, 88, 100, 100, 98**	**33, 33, 33, 67, 67**	**94, 98, 94, 94,93**

In general, despite the relatively high specificity values of type II and type III sequences and relatively high sensitivity values of type I sequences; sensitivity and specificity scores are not correlated in any way. These incompatibilities are probably due to low number of train and test sequences in type II and type III subclasses.

4.1.1.2 Gini Indexes

In our classification methods, each of the type I, type II and type III classification scores (CS) are calculated for sequences in the test set and the subclass with the highest CS is assigned to the relevant sequence. In an ideal case, a sequence would have the highest possible score for its own subclass and zero score for the others. However, sequences do not behave ideally in reality and a sequence has classification scores for each subclass.

Under these circumstances, the prediction will be as reliable as the degree of variance between classification scores. In order to express the variety between scores, Gini indexes are calculated for test sequences in each amino acid alphabet.

In the ideal case, sequences would have a zero Gini index. In our case, a smaller Gini Index indicates a better variety between subclass scores. The results are displayed in Table 4.2. These Gini Index values indicate slightly more than a two-fold overall variance.

The Gini index values display mild changes however they are in the same range for different RAAA and studies in general. Nevertheless, the Gini index changes do not show correlation for RAAA or for different studies.

4.1.1.3 Confidence Intervals

In addition to the importance of variance between classification scores, it is also crucial to determine how well the score of the assigned type represents its subclass. In order to define borders for scores, confidence intervals based on train set scores are calculated at 99% confidence level. Then for each alphabet and each dataset, the rate of the sequences with classification scores higher than the lower limit of the 99% confidence interval is calculated. Summarized in Table 4.4, the ratios are not very high; implying that nearly half of the test sequences do not significantly represent Type I population created by train set.

Table 4.4 The rate of the dockerin test sequences in 99% confidence intervals for all studies.

	Study1	Study2	Study3
	Rate of sequences in 99% confidence interval		
20-letter	51%	47%	70%
GMBR	65%	49%	52%
HSDM	36%	56%	58%
SDM	51%	43%	59%
Sezerman	51%	41%	62%

Cross-validation studies show some variety, but sequence rates do not display significantly high values. For amino acid alphabets, there is no correlation between different

studies. Thence, it is impossible to express whether any amino acid alphabet performs better in overall.

4.1.1.4 Profile-HMM Classification

In order to compare the proposed method with state-of-the-art classification methods, profile HMM classification is conducted on each dataset. Minimum error point test is applied to determine threshold values for HMM classification. Threshold values with error points and the accuracy rates for all datasets and all subclasses are denoted in Table 4.5.

Table 4.5 Profile HMM dockerin results for all subclasses and all studies are summarized. Minimum Error Point (MEP) is the threshold value used for HMM classification. FP and FN errors and the accuracy rate at that threshold level are shown.

		Study 1	Study 2	Study 3
Type I Prediction	MEP	7.17e-09	6.40e-05	9.20e-13
	FP	1	1	0
	FN	0	0	1
	Accuracy Rate	99%	99%	99%
Type II Prediction	MEP	x	2.03e+01	4.71e-20
	FP	x	17	0
	FN	x	0	0
	Accuracy Rate	x	76%	100%
Type III Prediction	MEP	8.02e-7	2.74e-05	2.85e-04
	FP	3	12	27
	FN	0	0	0
	Accuracy Rate	96%	83%	61%

In the case of type I prediction, HMM gives high accuracy rates and the accuracy rates on different studies are quite correlated. The accuracy rates of type II and type III classification show much more variance on different studies. This incompatibility is thought to arise from incompetent type II and type III motifs, as in the previous cases. The accuracy rates for Type I HMM prediction is higher than our method, however the accuracy rates of type II and type III HMM classification are not matching in different studies and give comparably low accuracy rates, as in the case of our proposed method. Specificity and sensitivity scores are calculated for HMM in Table 4.6. Sensitivity of HMM classification is generally higher than the sensitivity of the proposed method and

they are correlated for each study. However, our method gives higher specificity in general, compared to HMM.

Table 4.6 Dockerin sensitivity and specificity values of HMM. Values are calculated for prediction of each subclass on different studies.

	Type I		Type II		Type III	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Study 1	100	0	X	X	100	96
Study 2	100	50	100	75	100	83
Study 3	99	100	100	100	100	61

4.1.2 Subclass Identification for Cohesin

In this section, classification and performance results for cohesin sequences are discussed.

4.1.2.1 Confusion Matrix and Accuracy Rates

The confusion matrix of cohesin classification for each RAAA and each study is summarized in Table 4.7. All sequences are generally classified accurately, however in a few numbers of cases, type II and type III sequences are classified as type I. Unlikely, type II and type III sequences display no misclassification between them. In addition, the results are in correlation with each other in different RAAAs and studies.

The accuracy rates of the proposed classification method for studies on five different amino acid alphabets are listed in Table 4.8. It is noteworthy that the average classification accuracy for all amino acid alphabets is higher than 97%. The accuracy levels for different studies in RAAAs show correlation.

Table 4.7 The confusion matrix of cohesin classification. In each section, rows represent different RAAAs and columns represent the cases; study 1, study 2 and study 3, respectively.

Confusion Matrix	**Type I Actual**	**Type II Actual**	**Type III Actual**	**Alphabets**
Type I Predicted	36 36 36 36 36 36 36 36 35 36 36 35 36 36 36	0 0 0 0 0 0 0 1 1 0 0 1 0 0 0	0 1 1 0 1 1 0 1 1 0 1 1 0 1 1	20 letter GMBR HSDM SDM Sezerman
Type II Predicted	0 0 0 0 0 0 0 0 1 0 0 1 0 0 0	22 22 22 22 22 22 22 21 21 22 22 21 22 22 22	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	20 letter GMBR HSDM SDM Sezerman
Type III Predicted	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 4 4 5 4 4 5 4 4 5 4 4 5 4 4	20 letter GMBR HSDM SDM Sezerman

Table 4.8 The accuracy rates and Gini index of cohesin classification for different amino acid alphabets and for different studies.

Proposed Method	Accuracy Rate (%) Gini Index				
	20 letter	GMBR	HSDM	SDM	Sezerman
Study 1	100% *0.545*	100% *0.542*	100% *0.539*	100% *0.545*	100% *0.540*
Study 2	98% *0.497*	98% *0.613*	96% *0.617*	98% *0.549*	98% *0.558*
Study3	98% *0.535*	98% *0.598*	96% *0.589*	96% *0.579*	98% *0.546*
Average	*99%* *0.523*	*99%* *0.584*	*97%* *0.582*	*98%* *0.558*	*99%* *0.548*

Among the five different amino acids alphabets, despite the HSDM and SDM alphabets perform slightly worse than the other alphabets, the accuracy levels are not vastly different, as summarized in Table 4.8.

It is stated before that due to limited number of available type II and type III sequences, training sets of these subclasses are in small scale compared to type I train sets. Even though the number of available type II and type III cohesins are higher than dockerins, they still compose smaller portions compared to type I sequences. Thence, the question about the qualification of motifs to reflect the true conservation is also valid for cohesin sequences. Regardless of these problems, high classification accuracy levels are obtained for each amino acid alphabet using the proposed method. Following, the specificity and sensitivity levels of classification calculated from the confusion matrix are presented in Table 4.9.

Table 4.9 Cohesin sensitivity and specificity values calculated from confusion matrix for type I, type II and type III prediction on five different amino acid alphabets. Different colors represent different amino acid alphabets; 20-letter, GMBR, HSDM, SDM and Sezerman, respectively.

	Type I		**Type II**		**Type III**	
	Sensitivity (%)	**Specificity (%)**	**Sensitivity, (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
Study 1	**100,100,100, 100,100**	**100,100,100, 100,100**	**100,100,100, 100,100**	**100,100,100, 100,100**	**100,100,100, 100,100**	**100,100,100, 100,100**
Study 2	**100,100,100, 100,100**	**96, 96, 93 96, 96**	**100,100,95, 100,100**	**100,100,100, 100,100**	**80, 80, 80 80,80**	**100,100,100, 100,100**
Study 3	**100,100,100, 100,100**	**96, 96, 96 96, 96**	**100,100,100, 100,100**	**100,100,98, 98,100**	**80, 80, 80 80,80**	**100,100,100, 100,100**
Average	**100, 100, 99, 99, 100**	**97, 97, 96, 97, 97**	**100, 100, 98, 100, 100**	**100, 100, 99, 99, 100**	**87, 87, 87, 87, 87**	**100,100,100, 100,100**

In each study, on the contrary of the dockerin case, specificity and sensitivity scores are correlated. Specificity and sensitivity levels of HSDM and SDM alphabets show some variation in cross-validation studies, however, other alphabets produce the same values

for each case. In addition, sensitivity levels for prediction of type I and type II sequences are higher than the type III values. As stated in dockerin case, the relatively small dataset of type III sequences is the possible reason behind the low sensitivity level. However, the specificity levels for type III are comparably high, indicating the defined motifs display great specificity despite the small size dataset.

4.1.2.2 Gini Indexes

The Gini index results for cohesin classification are displayed in Table. The Gini index levels are smaller than dockerin values, indicating that the scoring introduced by our method is more discriminative in cohesin test. Gini index values show correlation between RAAA and studies, corresponding to a more or less two-fold difference.

4.1.2.3 Confidence Intervals

For each alphabet and each dataset, the rate of the sequences with classification scores higher than the lower limit of the 99% confidence interval is calculated and summarized in Table 4.10. Compared to dockerin sequences, a very high percentage of the cohesin test sequences obtain scores in 99% confidence interval. These values indicate that most test set sequence represent their assigned subclass population created by the train set. The cross-validation studies are correlated in general.

Table 4.10 The rate of the cohesin test sequences in 99% confidence intervals for all datasets.

	Study 1	Study 2	Study 3
	Rate of sequences in 99% confidence interval		
20-letter	85%	90%	89%
GMBR	90%	91%	90%
HSDM	87%	89%	90%
SDM	87%	89%	89%
Sezerman	87%	89%	89%

4.1.2.4 Profile-HMM Classification

Profile HMM analysis is conducted on cohesin test set for type I, type II and type III prediction. The results are summarized in Table 4.11. In each study, profile HMM gives correlated results for all RAAAs. Type I sequences are predicted with 100% accuracy, where type II and type III sequences are predicted with an average of 99% accuracy. HMM performs slightly better than HSDM and SDM alphabet in our method, however for the other alphabets, they give the same accuracy rate as HMM.

Table 4.11 Profile HMM cohesin results for all subclasses and all studies are summarized. Minimum Error Point (MEP) is the threshold value used for HMM classification. FP and FN errors and the accuracy rate at that threshold level are shown.

		Study 1	Study 2	Study 3
Type I Prediction	MEP	7.66e-08	4.58e-12	5.48e-08
	FP	0	0	0
	FN	0	0	0
	Accuracy Rate	100%	100%	100%
Type II Prediction	MEP	0.00e+00	0.00e+00	1.48e-266
	FP	0	0	0
	FN	2	0	0
	Accuracy Rate	97%	100%	100%
Type III Prediction	MEP	6.88e-02	5.98e-101	1.07e-74
	FP	0	0	0
	FN	0	1	1
	Accuracy Rate	100%	98%	98%

Table 4.12 Cohesin sensitivity and specificity values of HMM. Values are calculated for prediction of each subclass on different studies.

	Type I		Type II		Type III	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Study 1	100%	100%	91%	100%	100%	100%
Study 2	100%	100%	100%	100%	80%	100%
Study 3	100%	100%	100%	100%	80%	100%

Sensitivity and specificity for type I HMM prediction gives the same values as our method, except for the HSDM and SDM alphabets that performs slightly worse than the other alphabets. For type II prediction, our method displays sensitivity and specificity higher than 99% that outperforms the sensitivity of HMM. For type III sequences, HMM gives nearly the same scores as our prediction method (Table 4.12).

4.2 Classification of Sequences with Unknown Subclass

The classification results of dockerin and cohesin sequences with unknown subclass are provided in Table B.1. The accuracy of this classification demands to be determined by future studies that will define the subclasses for the below sequences.

4.3 Correlated Mutation Studies

Correlated mutation studies are conducted on type I, type II and type III datasets, separately. Type II and type III analysis do not express any correlated residues. We believe that the number of available type II and type III sequences is not enough to calculate evolutionary correlation. The correlated residues for type I sequences are summarized on Table 4.13. The residues highlighted in red demonstrate the positions at which the residues are defined as motifs in all alphabets and used for subclass prediction. In consistence with our intuition, correlated residues for dockerin-cohesin type I interaction show overlaps with the defined motifs in this study. In Table 4.14, it is explained in detail which motif is identified in which alphabet.

In this thesis, as well as classifying dockerin and cohesins into their subtypes, we aim to pinpoint possible key interaction sites. These highlighted residues are conserved among type I sequences and coevolve together. Thence, these residues can be defined as key site candidates for dockerin-cohesin interaction studies.

In Figure4.1, the residues that are corresponding to the correlated motifs are shown on a known type I Clostridium Cellulolyticum dockerin-cohesin complex structure. Dockerin residues are represented in red, whereas cohesin sequences are represented in yellow.

Table 4.13 Correlated dockerin-cohesin residues. Values indicate positions in aligned form, whereas the values in brackets display the residues in unaligned form. The residues highlighted in red are the residues correlated with motifs utilized in this study.

Dockerin Residues	Correlated Cohesin Residues
4(4)	22(20)
11(11)	82(68) (G)
12(12)	47(44) 50(47) 107(90) (T, S, N, Q) 117(100) (A, S, T) 124(103) 162(123)
13(13) (G)	3(1) 5(3)(F, I, L, Y) 16(14)(I, V) 18(16) 20(18)(I, V) 25(23)(I, V)
16(16)	20(18) (I, V) 25(23) (I, V) 117(100) (A, S, T)
17(17) (S, A, I)	120(101)
18(18)	21(19)
23(23)	107(90)(T, S, N, Q)
24(24) (R, K)	125(104) 149(111)
25(25) (Q, R, K)	117(100) (A, S, T)
27(27) (L, V, I)	3(1) 33(30) 149(111) 162(123)
31(31) (I, L, F)	38(35) (L, I, V, Y)
38(34) (F, L)	47(44) 76(63) 151(113)
42(38)	37(34) 150(112)
43(39)	76(63)
44(40)	71(58) 150(112) 151(113)
48(41)	150(112)
51(44) (A)	94(79) 107(90) (T, S, N, Q) 114(97) (K,S)
58(50) (G)	107(90) (T, S, N, Q) 120(101)
60(52) (V, F, I, A)	20(18)(I, V) 25(23)(I, V) 47(44) 94(79) 114(97)(K,S) 162(123)
62(54)	20(18)(I, V) 25(23)(I, V) 82(68)(G) 150(112) 151(113) 167(128)(V,I)

Table 4.14 The motifs overlapping with correlated sites and the alphabets that these motifs are defined. D stands for dockerin and C stands for cohesin residues.

Alphabets	Motifs defined in the corresponding alphabets
20-letter	C82, D13, D24, D51, D58
GMBR	C5, C16, C20, C25, C38, C114, C117, C167, D13, D17, D24, D58
HSDM	C16, C38,C82, C167, D13, D24, D27, D51, D58, D60
SDM	C5, C16, C20, C25, C38, C82, C107, C167, D24, D58, D60
Sezerman	C5, C16, C20, C25, C38, C82, C107, C167, D13, D24, D25, D27, D51, D58, D60

Figure 4.1 Representation of motifs that overlap with correlated sites on a known structure of type I Clostridium Cellulolyticum dockerin-cohesin complex (PDB code: 2VN6)

Chapter 5

CONCLUSIONS AND FUTURE PROSPECTS

In this thesis, we propose a method to classify dockerin and cohesin sequences into their subclasses using conserved amino acid residues as motifs. The basic difference between dockerin and cohesin subclasses is their mode of interaction. In this thesis, as well as accurate subclass prediction, we aim to define the key interaction site candidates for each subclass for design purposes.

In a multiple sequence alignment, even functionally or structurally critical amino acids can be substituted with physiochemical similar amino acids. Since we are working on conserved residues, we introduce reduced amino acid alphabets to catch these amino acids as motifs. We used four different reduced amino acid alphabets to analyze the effect of different groupings in our classification.

We performed our prediction using classification scores that are calculated based on motif specificities. For cohesin sequences, we obtain high classification accuracy, higher than 97%. For 20 letter, GMBR and Sezerman alphabets, the accuracy is even higher, 99%. Our method gives high sensitivity and specificity scores as well. 20 letter, GMBR and Sezerman alphabets gives the best sensitivity score, 100% and specificity score, 96% for type I and type II subclass prediction. For type III prediction, all alphabets give 80% sensitivity and 100%specificity scores.

For dockerins, on the other hand, the performance of the method drops. Accuracy rates are higher than 90% and 20 letter alphabet gives the best accuracy level with 95%; however, these scores are not correlated when different datasets are utilized. Sensitivity and specificity levels display drastic changes between different datasets and different amino acid alphabets. For HMM classification, the same pattern is observed. Type I sequences are classified with 91% accuracy; however this level drops to 61% for type III classification for example. On different datasets, the sensitivity and specificity levels

for HMM shows critical changes. Compared with high classification performance of cohesins, we consider the low number of type II and type III sequences as the source of this problem. This method has the potential for high performance classification as in the cohesin case, but for dockerin classification, it needs to be conducted on a dataset with more type II and type III sequences.

There are several widely-known protein classification methods; however these methods do not reveal any key residue information. In our method, we define visible motifs that are conserved in only the target subclass. Since these motifs are subclass specific, we claim that they have the potential to act as key interaction sites. Therefore, we performed correlated mutation studies between dockerin and cohesin sequences in order to approve our intuition. The motifs used in our method and coevolved residues of dockerins and cohesins show correlation, as expected. Thence, we propose the motifs overlapping with correlated mutation studies as key interaction site candidates.

BIBLIOGRAPHY

1. Jordan, D.B., et al., *Plant cell walls to ethanol.* Biochem J, 2012. **442**(2): p. 241-52.

2. Kroon-Batenburg, L.M. and J. Kroon, *The crystal and molecular structures of cellulose I and II.* Glycoconj J, 1997. **14**(5): p. 677-90.

3. Perez, J., et al., *Biodegradation and biological treatments of cellulose, hemicellulose and lignin: an overview.* Int Microbiol, 2002. **5**(2): p. 53-63.

4. Peterson, C.L. and T. Hustrulid, *Carbon cycle for rapeseed oil biodiesel fuels.* Biomass & Bioenergy, 1998. **14**(2): p. 91-101.

5. Chauvigne-Hines, L.M., et al., *Suite of Activity-Based Probes for Cellulose-Degrading Enzymes.* Journal of the American Chemical Society, 2012.

6. Das, H. and S.K. Singh, *Useful byproducts from cellulosic wastes of agriculture and food industry--a critical appraisal.* Crit Rev Food Sci Nutr, 2004. **44**(2): p. 77-89.

7. French, C.E., *Synthetic biology and biomass conversion: a match made in heaven?* J R Soc Interface, 2009. **6 Suppl 4**: p. S547-58.

8. Bayer, E.A., et al., *Cellulosomes-structure and ultrastructure.* J Struct Biol, 1998. **124**(2-3): p. 221-34.

9. Spinelli, S., et al., *Crystal structure of a cohesin module from Clostridium cellulolyticum: implications for dockerin recognition.* J Mol Biol, 2000. **304**(2): p. 189-200.

10. Noach, I., et al., *Modular arrangement of a cellulosomal scaffoldin subunit revealed from the crystal structure of a cohesin dyad.* J Mol Biol, 2010. **399**(2): p. 294-305.

11. Morais, S., et al., *Deconstruction of lignocellulose into soluble sugars by native and designer cellulosomes.* MBio, 2012. **3**(6).

12. Gefen, G., et al., *Enhanced cellulose degradation by targeted integration of a cohesin-fused beta-glucosidase into the Clostridium thermocellum cellulosome.* Proc Natl Acad Sci U S A, 2012. **109**(26): p. 10298-303.

13. Bayer, E.A., R. Lamed, and M.E. Himmel, *The potential of cellulases and cellulosomes for cellulosic waste management.* Curr Opin Biotechnol, 2007. **18**(3): p. 237-45.

14. Peer, A., et al., *Noncellulosomal cohesin- and dockerin-like modules in the three domains of life.* FEMS Microbiol Lett, 2009. **291**(1): p. 1-16.

15. Rincon, M.T., et al., *Unconventional mode of attachment of the Ruminococcus flavefaciens cellulosome to the cell surface.* J Bacteriol, 2005. **187**(22): p. 7569-78.

16. Adams, J.J., et al., *Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex.* Proc Natl Acad Sci U S A, 2006. **103**(2): p. 305-10.

17. Lytle, B.L., et al., *Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain.* J Mol Biol, 2001. **307**(3): p. 745-53.

18. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. **157**(1): p. 105-32.

19. Lengyel, P., *Problems in protein biosynthesis.* J Gen Physiol, 1966. **49**(6): p. 305-30.

20. Ptitsyn, O.B., *[Stages in the mechanism of self-organization of protein molecules].* Dokl Akad Nauk SSSR, 1973. **210**(5): p. 1213-5.

21. Fersht, A.R., *From the first protein structures to our current knowledge of protein folding: delights and scepticisms.* Nature Reviews Molecular Cell Biology, 2008. **9**(8): p. 650-654.

22. Dill, K.A. and J.L. MacCallum, *The protein-folding problem, 50 years on.* Science, 2012. **338**(6110): p. 1042-6.

23. Doi, R.H. and A. Kosugi, *Cellulosomes: plant-cell-wall-degrading enzyme complexes.* Nat Rev Microbiol, 2004. **2**(7): p. 541-51.

24. Burton, R.A., M.J. Gidley, and G.B. Fincher, *Heterogeneity in the chemistry, structure and function of plant cell walls.* Nat Chem Biol, 2010. **6**(10): p. 724-32.

25. Bayer, E.A., et al., *Cellulose, cellulases and cellulosomes.* Curr Opin Struct Biol, 1998. **8**(5): p. 548-57.

26. Fernandes, A.N., et al., *Nanostructure of cellulose microfibrils in spruce wood.* Proc Natl Acad Sci U S A, 2011. **108**(47): p. E1195-203.

27. Larsson, P.T., et al., *CP/MAS 13C-NMR spectroscopy applied to structure and interaction studies on cellulose I.* Solid State Nucl Magn Reson, 1999. **15**(1): p. 31-40.

28. Gibson, L.J., *The hierarchical structure and mechanics of plant materials.* J R Soc Interface, 2012. **9**(76): p. 2749-66.

29. Doi, R.H. and Y. Tamaru, *The Clostridium cellulovorans cellulosome: an enzyme complex with plant cell wall degrading activity.* Chem Rec, 2001. **1**(1): p. 24-32.

30. Ding, S.Y., et al., *How does plant cell wall nanoscale architecture correlate with enzymatic digestibility?* Science, 2012. **338**(6110): p. 1055-60.

31. van Driesser, B. and L. Christopher, *Mechanisms prevalent during bioremediation of wastewaters from the pulp and paper industry.* Crit Rev Biotechnol, 2004. **24**(2-3): p. 85-95.

32. Hipler, U.C., P. Elsner, and J.W. Fluhr, *A new silver-loaded cellulosic fiber with antifungal and antibacterial properties.* Curr Probl Dermatol, 2006. **33**: p. 165-78.

33. Gilbert, H.J., *Cellulosomes: microbial nanomachines that display plasticity in quaternary structure.* Mol Microbiol, 2007. **63**(6): p. 1568-76.

34. Xu, J. and J.C. Smith, *Probing the mechanism of cellulosome attachment to the Clostridium thermocellum cell surface: computer simulation of the Type II cohesin-dockerin complex and its variants.* Protein Eng Des Sel, 2010. **23**(10): p. 759-68.

35. Haimovitz, R., et al., *Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules.* Proteomics, 2008. **8**(5): p. 968-79.

36. Adams, J.J., et al., *Structural characterization of type II dockerin module from the cellulosome of Clostridium thermocellum: calcium-induced effects on conformation and target recognition.* Biochemistry, 2005. **44**(6): p. 2173-82.

37. Noach, I., et al., *Crystal structure of a type-II cohesin module from the Bacteroides cellulosolvens cellulosome reveals novel and distinctive secondary structural elements.* J Mol Biol, 2005. **348**(1): p. 1-12.

38. Xu, Q., et al., *The cellulosome system of Acetivibrio cellulolyticus includes a novel type of adaptor protein and a cell surface anchoring protein.* J Bacteriol, 2003. **185**(15): p. 4548-57.

39. Ding, S.Y., et al., *A scaffoldin of the Bacteroides cellulosolvens cellulosome that contains 11 type II cohesins.* J Bacteriol, 2000. **182**(17): p. 4915-25.

40. Xu, Q., et al., *Architecture of the Bacteroides cellulosolvens cellulosome: description of a cell surface-anchoring scaffoldin and a family 48 cellulase.* J Bacteriol, 2004. **186**(4): p. 968-77.

41. Rincon, M.T., et al., *ScaC, an adaptor protein carrying a novel cohesin that expands the dockerin-binding repertoire of the Ruminococcus flavefaciens 17 cellulosome.* J Bacteriol, 2004. **186**(9): p. 2576-85.

42. Shimon, L.J., et al., *A cohesin domain from Clostridium thermocellum: the crystal structure provides new insights into cellulosome assembly.* Structure, 1997. **5**(3): p. 381-90.

43. Pinheiro, B.A., et al., *Functional insights into the role of novel type I cohesin and dockerin domains from Clostridium thermocellum.* Biochemical Journal, 2009. **424**: p. 375-384.

44. Pinheiro, B.A., et al., *The Clostridium cellulolyticum dockerin displays a dual binding mode for its cohesin partner.* Journal of Biological Chemistry, 2008. **283**(26): p. 18422-30.

45. Carvalho, A.L., et al., *Evidence for a dual binding mode of dockerin modules to cohesins.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3089-94.

46. Ding, S.Y., et al., *Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens.* J Bacteriol, 2001. **183**(6): p. 1945-53.

47. Karpol, A., et al., *Structural and functional characterization of a novel type-III dockerin from Ruminococcus flavefaciens.* FEBS Lett, 2012.

48. Bayer, E.A., P.M. Coutinho, and B. Henrissat, *Cellulosome-like sequences in Archaeoglobus fulgidus: an enigmatic vestige of cohesin and dockerin domains.* FEBS Lett, 1999. **463**(3): p. 277-80.

49. Voronov-Goldman, M., et al., *Noncellulosomal cohesin from the hyperthermophilic archaeon Archaeoglobus fulgidus.* Proteins, 2011. **79**(1): p. 50-60.

50. Lamed, R., E. Setter, and E.A. Bayer, *Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum.* J Bacteriol, 1983. **156**(2): p. 828-36.

51. Desvaux, M., et al., *Genomic analysis of the protein secretion systems in Clostridium acetobutylicum ATCC 824.* Biochim Biophys Acta, 2005. **1745**(2): p. 223-53.

52. *The clostridia and biotechnology.* Biotechnology, 1993. **25**: p. 1-429.

53. Morisaka, H., et al., *Profile of native cellulosomal proteins of Clostridium cellulovorans adapted to various carbon sources.* AMB Express, 2012. **2**(1): p. 37.

54. Blouzard, J.C., et al., *Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses.* Proteomics, 2010. **10**(3): p. 541-54.

55. Perret, S., et al., *Towards designer cellulosomes in Clostridia: mannanase enrichment of the cellulosomes produced by Clostridium cellulolyticum.* J Bacteriol, 2004. **186**(19): p. 6544-52.

56. Kamezaki, Y., et al., *The Dock tag, an affinity tool for the purification of recombinant proteins, based on the interaction between dockerin and cohesin domains from Clostridium josui cellulosome.* Protein Expr Purif, 2010. **70**(1): p. 23-31.

57. Jindou, S., et al., *Cohesin-dockerin interactions within and between Clostridium josui and Clostridium thermocellum: binding selectivity between cognate dockerin and cohesin domains and species specificity.* Journal of Biological Chemistry, 2004. **279**(11): p. 9867-74.

58. Leibovitz, E. and P. Beguin, *A new type of cohesin domain that specifically binds the dockerin domain of the Clostridium thermocellum cellulosome-integrating protein CipA.* J Bacteriol, 1996. **178**(11): p. 3077-84.

59. Dassa, B., et al., *Genome-wide analysis of acetivibrio cellulolyticus provides a blueprint of an elaborate cellulosome system.* BMC Genomics, 2012. **13**: p. 210.

60. Xu, Q., et al., *A novel Acetivibrio cellulolyticus anchoring scaffoldin that bears divergent cohesins.* J Bacteriol, 2004. **186**(17): p. 5782-9.

61. Jindou, S., et al., *Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium Ruminococcus flavefaciens.* FEMS Microbiol Lett, 2008. **285**(2): p. 188-94.

62. Rincon, M.T., et al., *A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of Ruminococcus flavefaciens.* J Bacteriol, 2007. **189**(13): p. 4774-83.

63. Ruepp, A., et al., *The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(18): p. 5539-45.

64. Wu, C.H., et al., *Protein family classification and functional annotation.* Comput Biol Chem, 2003. **27**(1): p. 37-47.

65. Brown, D.P., N. Krishnamurthy, and K. Sjolander, *Automated protein subfamily identification and classification.* PLoS Comput Biol, 2007. **3**(8): p. e160.

66. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.

67. Cai, C.Z., et al., *Protein function classification via support vector machine approach.* Math Biosci, 2003. **185**(2): p. 111-22.

68. Cobanoglu, M.C., Y. Saygin, and U. Sezerman, *Classification of GPCRs using family specific motifs.* IEEE/ACM Trans Comput Biol Bioinform, 2011. **8**(6): p. 1495-508.

69. Wistrand, M. and E.L. Sonnhammer, *Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER.* BMC Bioinformatics, 2005. **6**: p. 99.

70. Schliep, A., A. Schonhuth, and C. Steinhoff, *Using hidden Markov models to analyze gene expression time course data.* Bioinformatics, 2003. **19**: p. i255-i263.

71. Lyngso, R.B., C.N. Pedersen, and H. Nielsen, *Metrics and similarity measures for hidden Markov models.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 178-86.

72. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.

73. Krogh, A., et al., *Hidden Markov models in computational biology. Applications to protein modeling.* J Mol Biol, 1994. **235**(5): p. 1501-31.

74. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.

75. Noble, W.S., *What is a support vector machine?* Nature Biotechnology, 2006. **24**(12): p. 1565-1567.

76. O'Dwyer, L., et al., *Using Support Vector Machines with Multiple Indices of Diffusion for Automated Classification of Mild Cognitive Impairment.* PLoS One, 2012. **7**(2).

77. Brown, M.P.S., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(1): p. 262-267.

78. Cai, Y.D., et al., *Support vector machines for the classification and prediction of beta-turn types.* J Pept Sci, 2002. **8**(7): p. 297-301.

79. Cai, C.Z., et al., *Protein function classification via support vector machine approach.* Mathematical Biosciences, 2003. **185**(2): p. 111-122.

80. Ding, C.H.Q. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks.* Bioinformatics, 2001. **17**(4): p. 349-358.

81. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure.* Bioinformatics, 2001. **17**(5): p. 455-60.

82. Ouzounis, C.A., et al., *Classification schemes for protein structure and function.* Nat Rev Genet, 2003. **4**(7): p. 508-19.

83. Krogh, A., et al., *Hidden Markov-Models in Computational Biology - Applications to Protein Modeling.* Journal of Molecular Biology, 1994. **235**(5): p. 1501-1531.

84. Karchin, R., K. Karplus, and D. Haussler, *Classifying G-protein coupled receptors with support vector machines.* Bioinformatics, 2002. **18**(1): p. 147-59.

85. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

86. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

87. Weston, J., et al., *Semi-supervised protein classification using cluster kernels.* Bioinformatics, 2005. **21**(15): p. 3241-3247.

88. Jaakkola, T., M. Diekhans, and D. Haussler, *A discriminative framework for detecting remote protein homologies.* Journal of Computational Biology, 2000. **7**(1-2): p. 95-114.

89. Henikoff, S. and J.G. Henikoff, *Performance evaluation of amino acid substitution matrices.* Proteins, 1993. **17**(1): p. 49-61.

90. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

91. Prlic, A., F.S. Domingues, and M.J. Sippl, *Structure-derived substitution matrices for alignment of distantly related sequences.* Protein Eng, 2000. **13**(8): p. 545-50.

92. Peterson, E.L., et al., *Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment.* Bioinformatics, 2009. **25**(11): p. 1356-1362.

93. Solis, A.D. and S. Rackovsky, *Optimized representations and maximal information in proteins.* Proteins, 2000. **38**(2): p. 149-64.

94. Erguner, B., O. Erdogan, and U. Sezerman, *Prediction and classification for GPCR sequences based on ligand specific features.* Computer and Information Sciences - ISCIS 2006, Proceedings, 2006. **4263**: p. 174-181.

95. Dill, K.A., *Theory for the folding and stability of globular proteins.* Biochemistry, 1985. **24**(6): p. 1501-9.

96. Li, H., et al., *Emergence of preferred structures in a simple model of protein folding.* Science, 1996. **273**(5275): p. 666-9.

97. Hecht, M.H., et al., *De novo proteins from designed combinatorial libraries.* Protein Sci, 2004. **13**(7): p. 1711-23.

98. Hill, R.E. and N.D. Hastie, *Accelerated evolution in the reactive centre regions of serine protease inhibitors.* Nature, 1987. **326**(6108): p. 96-9.

99. Fraser, H.B., et al., *Evolutionary rate in the protein interaction network.* Science, 2002. **296**(5568): p. 750-2.

100. Pollock, D.D. and W.R. Taylor, *Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution.* Protein Engineering, 1997. **10**(6): p. 647-657.

101. Neher, E., *How frequent are correlated changes in families of protein sequences?* Proc Natl Acad Sci U S A, 1994. **91**(1): p. 98-102.

102. Halperin, I., H. Wolfson, and R. Nussinov, *Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.* Proteins, 2006. **63**(4): p. 832-45.

103. Gobel, U., et al., *Correlated mutations and residue contacts in proteins.* Proteins, 1994. **18**(4): p. 309-17.

104. Kass, I. and A. Horovitz, *Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations.* Proteins, 2002. **48**(4): p. 611-7.

105. Atchley, W.R., et al., *Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.* Mol Biol Evol, 2000. **17**(1): p. 164-78.

106. Galitsky, B., *Revealing the set of mutually correlated positions for the protein families of immunoglobulin fold.* In Silico Biol, 2003. **3**(3): p. 241-64.

107. Dekker, J.P., et al., *A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.* Bioinformatics, 2004. **20**(10): p. 1565-72.

108. Pollock, D.D., W.R. Taylor, and N. Goldman, *Coevolving protein residues: maximum likelihood identification and relationship to structure.* J Mol Biol, 1999. **287**(1): p. 187-98.

109. Schneider, M., T.U. Consortium, and S. Poux, *UniProtKB amid the turmoil of plant proteomics research.* Front Plant Sci, 2012. **3**: p. 270.

110. Ouzounis, C., et al., *Are binding residues conserved?* Pac Symp Biocomput, 1998: p. 401-12.

111. Reddy, B.V., et al., *Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins.* Proteins, 2001. **42**(2): p. 148-63.

112. Schein, C.H., O. Ivanciuc, and W. Braun, *Bioinformatics approaches to classifying allergens and predicting cross-reactivity.* Immunol Allergy Clin North Am, 2007. **27**(1): p. 1-27.

113. Guharoy, M. and P. Chakrabarti, *Conserved residue clusters at protein-protein interfaces and their use in binding site identification.* BMC Bioinformatics, 2010. **11**.

114. Sjolander, K., et al., *Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.* Comput Appl Biosci, 1996. **12**(4): p. 327-45.

115. Lee, W.C., *Characterizing exposure-disease association in human populations using the Lorenz curve and Gini index.* Stat Med, 1997. **16**(7): p. 729-39.

116. Qi, Y.J., Z. Bar-Joseph, and J. Klein-Seetharaman, *Evaluation of different biological data and computational classification methods for use in protein interaction prediction.* Proteins-Structure Function and Bioinformatics, 2006. **63**(3): p. 490-500.

117. Dunn, O.J. and V. Clark, *Applied statistics : analysis of variance and regression*. 2nd ed. Wiley series in probability and mathematical statistics Applied probability and statistics. 1987, New York: Wiley. xii, 445 p.

118. Fares, M.A. and D. McNally, *CAPS: coevolution analysis using protein sequences.* Bioinformatics, 2006. **22**(22): p. 2821-2822.

Appendix A

Motif, Positions and Motif Specificity Scores

Table A.1 Motifs used in cohesin 20-letter alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			31	A	75	164	V	71
16	V	79	33	Y	92	166	Q	75
19	P	95	34	Q	100	164	V	71
36	F	87	36	N	92	166	Q	75
40	Y	92	38	K	100	166	Q	75
41	D	97	39	Y	88	**Type III**		
45	L	89	40	D	96	7	W	100
53	G	100	41	P	96	16	P	67
73	F	87	52	G	88	17	G	100
82	G	71	61	P	96	37	A	83
87	L	89	65	G	92	38	G	83
88	F	97	74	Y	96	40	Q	100
100	I	100	77	T	75	41	F	100
105	V	82	91	F	88	51	Y	100
106	F	95	95	Y	100	60	Y	100
111	F	76	98	L	83	61	G	83
112	K	87	105	G	83	75	K	100
165	G	95	111	G	100	78	F	83
Type II			116	I	75	80	F	100
8	D	96	118	F	100	103	V	100
10	T	83	120	V	92	110	G	100
15	G	100	146	G	100	113	Y	100
16	D	88	152	W	100	132	V	67
30	F	71	154	G	88			

Table A.2 Motifs used in cohesin GMBR alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			**Type II**			154	G	88
1	Y	82	15	G	100	164	Y	75
7	A	95	30	Y	100	**Type III**		
16	Y	95	33	Y	100	16	P	67
19	P	95	36	A	96	17	G	100
31	Y	97	41	P	96	18	A	100
38	Y	97	52	G	88	38	G	83
40	Y	100	61	P	96	41	Y	100
45	Y	100	65	G	92	61	G	83
53	G	100	66	A	96	98	A	67
56	Y	100	74	Y	100	103	Y	100
73	Y	100	104	A	92	110	G	100
82	G	71	105	G	83	117	Y	100
87	Y	89	111	G	100	132	Y	67
101	A	97	120	Y	92	134	A	67
105	Y	95	132	A	100	135	A	67
111	Y	76	137	A	79			
164	A	95	146	G	100			
165	G	95	148	Y	88			
167	Y	84	152	Y	100			

Table A.3 Motifs used in cohesin HSDM alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			**Type II**			146	G	100
16	L	95	7	L	88	149	L	71
19	P	95	8	D	96	152	W	100
31	L	95	10	T	83	154	G	88
36	F	87	15	G	100	164	L	75
38	L	79	16	D	88	166	Q	75
40	Y	92	30	F	71	**Type III**		
41	D	97	31	A	75	7	W	100
45	L	95	33	Y	92	16	P	67
48	L	76	34	Q	100	17	G	100
53	G	100	36	N	92	36	L	100
56	L	100	38	E	100	37	A	83
73	F	87	39	Y	88	38	G	83
82	G	71	40	D	96	40	Q	100
87	L	89	41	P	96	41	F	100
88	F	97	52	G	88	51	Y	100
105	L	95	61	P	96	60	Y	100
106	F	95	65	G	92	61	G	83
109	L	100	74	Y	96	75	E	100
111	F	76	77	T	75	78	F	83
112	E	87	91	F	88	80	F	100
165	G	95	95	Y	100	103	L	100
167	L	84	105	G	83	110	G	100
			111	G	100	113	Y	100
			118	F	100	132	L	67
			120	L	92			

Table A.4 Motifs used in cohesin SDM alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			**Type II**			118	Y	100
1	L	74	7	L	88	120	L	92
16	L	95	8	D	96	139	T	71
19	P	95	10	T	88	146	G	100
31	L	97	15	G	100	149	L	100
36	Y	87	16	D	88	152	W	100
38	L	79	30	Y	71	154	G	88
40	Y	100	31	A	75	164	L	75
41	D	97	33	Y	92	**Type III**		
45	L	97	36	N	92	7	W	100
48	L	76	38	E	100	16	P	67
53	G	100	39	Y	100	17	G	100
56	L	100	40	D	96	37	A	83
73	Y	92	41	P	96	38	G	83
82	G	71	52	G	88	40	T	100
87	L	89	55	Y	71	41	Y	100
88	Y	100	59	T	83	51	Y	100
105	L	95	60	L	71	60	Y	100
106	Y	95	61	P	96	61	G	83
109	L	100	65	G	92	75	E	100
111	Y	76	74	Y	100	80	Y	100
112	E	87	95	Y	100	103	L	100
165	G	95	104	T	79	110	G	100
167	L	84	105	G	83	113	Y	100
			111	G	100	132	L	67

Table A.5 Motifs used in cohesin Sezerman alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			**Type II**			118	Y	100
1	L	74	7	L	88	120	L	92
16	L	95	8	D	96	139	T	71
19	P	95	10	T	88	145	D	75
31	L	97	15	G	100	146	G	100
36	Y	87	16	D	96	149	L	100
38	L	79	30	Y	71	152	W	100
40	Y	100	31	A	75	154	G	88
41	D	97	33	Y	92	164	L	75
45	L	97	34	Q	100	166	Q	75
48	L	76	36	Q	92	**Type III**		
53	G	100	38	K	100	7	W	100
56	L	100	39	Y	100	16	P	67
73	Y	92	40	D	96	17	G	100
82	G	71	41	P	96	37	A	83
87	L	89	52	G	88	38	G	83
88	Y	100	55	Y	71	41	Y	100
90	D	92	59	T	83	51	Y	100
105	L	95	60	L	71	60	Y	100
106	Y	95	61	P	96	61	G	83
109	L	100	65	G	92	80	Y	100
111	Y	76	74	Y	100	103	L	100
112	K	87	95	Y	100	106	D	67
165	G	95	104	T	79	110	G	100
167	L	84	105	G	83	113	Y	100
			111	G	100	132	L	67

Table A.6 Motifs used in dockerin 20-letter alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			73	L	78	55	I	100
24	K	83	76	I	73	60	F	98
28	L	71	**Type II**			**Type III**		
54	N	71	27	F	100	45	T	100
58	G	78	41	D	92	48	G	100
64	D	100	46	G	100	67	D	100
68	L	75	48	I	95			

Table A.7 Motifs used in dockerin GMBR alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			**Type II**			59	Y	95
28	Y	92	17	Y	91	**Type III**		
58	G	78	36	Y	98	28	A	91
71	Y	90	46	G	100	46	A	94
72	Y	98	50	Y	94	48	G	100
76	Y	78	54	Y	94	57	A	95
79	Y	73	56	Y	98			

Table A.8 Motifs used in dockerin HSDM alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			72	L	97	56	L	100
24	E	83	73	L	92	60	F	98
28	L	86	76	L	78	**Type III**		
54	N	71	**Type II**			45	T	100
58	G	78	27	F	100	48	G	100
60	L	93	41	D	92	57	E	100
64	D	100	46	G	100	67	D	100

Table A.9 Motifs used in dockerin SDM alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			71	Y	78	54	L	100
24	E	93	72	L	98	56	L	100
28	L	92	73	L	93	60	Y	98
54	N	71	76	L	78	**Type III**		
58	G	78	**Type II**			45	T	98
60	L	93	17	L	90	48	G	100
64	D	100	27	Y	98	57	E	100
67	L	71	41	D	92	67	D	100
69	E	76	46	G	100			
70	E	78	50	L	94			

Table A.10 Motifs used in dockerin Sezerman alphabet classification with positions and MSSs.

Position	Motif	MSS	Position	Motif	MSS	Position	Motif	MSS
Type I			70	K	76	50	L	94
24	K	93	71	Y	78	54	L	100
28	L	92	72	L	98	56	L	100
54	Q	71	73	L	93	60	Y	98
58	G	78	76	L	78	**Type III**		
60	L	93	**Type II**			45	T	100
64	D	100	17	L	91	46	D	100
67	L	71	27	Y	98	48	G	100
69	K	83	46	G	100	67	D	98

Appendix B

Classification Results of Sequences with Unknown Subclass

Table B.1 The classification results of dockerin and cohesin sequences with unknown subclass utilizing the method proposed in the thesis.

sequences	20-letter	GMBR	HSDM	SDM	Sezerman
Ac303238224	Type I	type I	Type I	Type I	Type I
Ac303238225	Type I	type I	Type I	Type I	Type I
Ac303238226	Type I	type I	Type I	Type I	Type I
Ac303238253	Type III	type I	Type I	Type I	Type III
Ac303238258	Type I	type I	Type I	Type I	Type I
Ac303238264	Type I	type I	Type I	Type I	Type I
Ac303238279	Type I	type I	Type I	Type I	Type I
Ac303238386	Type I	type I	Type I	Type I	Type I
Ac303238400	Type I	type I	Type II	Type I	Type I
Ac303238468	Type I	type I	Type I	Type I	Type I
Ac303238547	Type I	type I	Type I	Type I	Type I
Ac303238632	Type I	type I	Type I	Type I	Type I
Ac303238713	Type I	type I	Type I	Type I	Type I
Ac303238767	Type I	type I	Type I	Type I	Type I
Ac303238768	Type I	type I	Type I	Type I	Type I
Ac303238773	Type I	type I	Type I	Type I	Type I
Ac303238777	Type I	type I	Type I	Type I	Type III
Ac303238778	Type I	type I	Type I	Type I	Type I
Ac303238897	Type II	type II	Type II	Type II	Type II
Ac303238922	Type II	type II	Type II	Type II	Type II
Ac303238957	Type I	type I	Type I	Type I	Type I
Ac303238961	Type I	type I	Type I	Type I	Type I
Ac303238962	Type I	type I	Type I	Type I	Type I
Ac303238963	Type I	type I	Type I	Type I	Type I
Ac303238964	Type I	type I	Type I	Type I	Type I
Ac303238965	Type I	type I	Type I	Type I	Type I
Ac303238968	Type III	type I	Type I	Type I	Type I
Ac303238981	Type I	type I	Type I	Type I	Type I
Ac303239140	Type I	type I	Type I	Type I	Type I
Ac303239155	Type I	type I	Type I	Type I	Type I
Ac303239557	Type I	type I	Type I	Type I	Type I
Ac303239704	Type I	type I	Type I	Type I	Type III
Ac303239720	Type III	type I	Type I	Type I	Type I
Ac303239731	Type I	type II	Type II	Type II	Type II

Ac303239811	Type III	type I	Type I	Type I	Type I
Ac303239861	Type I	type I	Type I	Type I	Type I
Ac303239871	Type III	type I	Type I	Type I	Type I
Ac303239892	Type II	type I	Type I	Type I	Type I
Ac303239977	Type I	type I	Type I	Type I	Type I
Ac303240010	Type I	type I	Type I	Type I	Type I
Ac303240017	Type III	type I	Type I	Type I	Type I
Ac303240086	Type I	type I	Type I	Type I	Type I
Ac303240301	Type III	type I	Type I	Type I	Type III
Ac303240314	Type I	type I	Type I	Type I	Type I
Ac303240334	Type I	type I	Type I	Type I	Type I
Ac303240350	Type I	type I	Type I	Type I	Type I
Ac303240387	Type I	type I	Type I	Type I	Type I
Ac303240398	Type I	type I	Type I	Type I	Type I
Ac303240529	Type I	type I	Type I	Type I	Type I
Ac303240580	Type II	type I	Type I	Type I	Type I
Ac303240605	Type I	type I	Type I	Type I	Type I
Ac303240606	Type II	type II	Type II	Type II	Type II
Ac303240624	Type I	type I	Type I	Type I	Type I
Ac303240716	Type III	type I	Type I	Type I	Type III
Ac303240869	Type I	type I	Type I	Type I	Type I
Ac303240877	Type III	type I	Type I	Type I	Type III
Ac303241008	Type I	type I	Type I	Type I	Type I
Ac303241016	Type III	type I	Type I	Type I	Type I
Ac303241026	Type I	type I	Type I	Type I	Type I
Ac303241027	Type I	type I	Type I	Type I	Type I
Ac303241061	Type I	type I	Type I	Type I	Type I
Ac303241098	Type I	type I	Type I	Type I	Type I
Ac303241149	Type I	type I	Type I	Type I	Type I
Ac303241211	Type III	type I	Type I	Type I	Type I
Ac303241235	Type I	type I	Type I	Type I	Type I
Ac303241236	Type I	type I	Type I	Type I	Type I
Ac303241237	Type III	type I	Type I	Type I	Type III
Ac303241300	Type I	type I	Type I	Type I	Type I
Ac303241524	Type I	type I	Type I	Type I	Type I
Ac303241536	Type I	type I	Type I	Type I	Type I
Ac303241822	Type I	type I	Type I	Type I	Type I
Ac303241877	Type I	type I	Type I	Type I	Type I
Ac303241878	Type I	type I	Type I	Type I	Type I
Ac303241889	Type I	type I	Type I	Type I	Type I
Ac303242281	Type I	type I	Type I	Type I	Type I
Ac303242294	Type I	type I	Type I	Type I	Type I
Ac303242527	Type III	type I	Type I	Type I	Type I
Ac303242528	Type III	type I	Type I	Type I	Type I
Ac303242586	Type I	type I	Type I	Type I	Type I
Ac303242589	Type I	type I	Type I	Type I	Type I
Ac303242732	Type I	type I	Type I	Type I	Type I
Ac303242895	Type I	type I	Type I	Type I	Type I
Ac303242911	Type I	type I	Type I	Type I	Type III
Ac303242986	Type I	type I	Type I	Type I	Type I

Ac303243005	Type I	type I	Type I	Type I	Type I
Ac303243136	Type I	type I	Type I	Type I	Type I
Ac31540575_gb_A	Type I	type I	Type I	Type I	Type II
Ac6249561_gb_AA	Type II	type II	Type II	Type II	Type II
Ac63252967_emb_	Type I	type I	Type I	Type I	Type I
c2782_256753117	Type II	type I	Type I	Type I	Type I
c2782_256753118	Type I	type I	Type I	Type I	Type I
c2782_256753123	Type I	type I	Type I	Type I	Type I
c2782_256753214	Type I	type I	Type I	Type I	Type I
c2782_256753275	Type I	type I	Type I	Type I	Type I
c2782_256753311	Type I	type I	Type I	Type I	Type I
c2782_256753349	Type I	type I	Type I	Type I	Type I
c2782_256753641	Type I	type I	Type I	Type I	Type I
c2782_256754017	Type I	type I	Type I	Type I	Type I
c2782_256754290	Type I	type I	Type I	Type I	Type I
c2782_256754672	Type I	type I	Type I	Type I	Type I
c2782_256754777	Type I	type I	Type I	Type I	Type I
c2782_256754780	Type I	type II	Type I	Type I	Type I
c2782_256754925	Type I	type I	Type I	Type I	Type I
c2782_256755005	Type I	type I	Type I	Type I	Type I
c2782_256755009	Type I	type I	Type I	Type I	Type I
c2782_256755010	Type I	type I	Type I	Type I	Type I
c2782_256755015	Type I	type I	Type I	Type I	Type I
c2782_256755016	Type I	type I	Type I	Type I	Type I
c2782_256755017	Type I	type I	Type I	Type I	Type I
c2782_256755018	Type I	type I	Type I	Type I	Type I
c2782_256755193	Type I	type I	Type I	Type I	Type I
c2782_256755328	Type I	type I	Type I	Type I	Type I
c2782_256755329	Type I	type I	Type I	Type I	Type I
c2782_256755554	Type I	type I	Type I	Type I	Type I
c2782_256755559	Type I	type I	Type I	Type I	Type I
c2782_256755593	Type I	type I	Type I	Type I	Type I
c2782_256755594	Type I	type I	Type I	Type I	Type I
c2782_256755643	Type I	type I	Type I	Type I	Type I
c2782_256755664	Type I	type I	Type I	Type I	Type I
c2782_256755969	Type I	type I	Type I	Type I	Type I
c2782_256756491	Type I	type I	Type I	Type I	Type I
c2782_256756492	Type I	type I	Type I	Type I	Type I
c2782_256756493	Type I	type I	Type I	Type I	Type I
c2782_256756496	Type I	type I	Type I	Type I	Type I
c2782_256756503	Type I	type I	Type I	Type I	Type I
c2782_256756505	Type I	type I	Type I	Type I	Type I
c2782_256756507	Type I	type I	Type I	Type I	Type I
c2782_256756508	Type I	type I	Type I	Type I	Type I
c2782_256756510	Type I	type I	Type I	Type I	Type I
c2782_256756511	Type I	type I	Type I	Type I	Type I
c2782_256756512	Type I	type I	Type I	Type I	Type I
c2782_256756513	Type I	type I	Type I	Type I	Type I
c2782_256756563	Type I	type I	Type I	Type I	Type I
c2782_256756563	Type I	type I	Type I	Type I	Type I

c2782_256756949	Type I	type I	Type I	Type I	Type I
c2782_256757051	Type I	type I	Type I	Type I	Type I
c2782_256757052	Type I	type I	Type I	Type I	Type I
c2782_256757053	Type I	type I	Type I	Type I	Type I
c2782_256757054	Type I	type I	Type I	Type I	Type I
c2782_256757068	Type I	type I	Type I	Type I	Type I
c2782_256757078	Type I	type I	Type I	Type I	Type I
c2782_256757159	Type I	type I	Type I	Type I	Type I
c2782_256757445	Type I	type I	Type I	Type I	Type I
c2782_256757446	Type I	type I	Type I	Type I	Type I
c2782_256757448	Type I	type I	Type I	Type I	Type I
c2782_256757449	Type I	type I	Type I	Type I	Type I