NAGOYA UNIVERSITY
DOCTORAL THESIS

A study of Protean Segments (ProSs):- Short regions in Intrinsically Disordered Proteins (IDPs) that undergo disorder-to-order transitions upon binding

Author:
Divya Shaji

Supervisor:
Prof. Motonori Ota

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the

Ota Laboratory
Graduate School of Information Science

Declaration of Authorship

I, Divya Shaji, declare that this thesis titled, “A study of Protean Segments (ProSs):- Short regions in Intrinsically Disordered Proteins(IDPs) that undergo disorder-to- order transitions upon binding” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a PhD at this University.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:

Date:

Preface

This dissertation describes my work done in the Doctoral Program of Bioinformatics in the Ota Laboratory, Department of Complex Systems Science, Graduate School of Information Science, Nagoya University, under the supervision of Prof. Motonori Ota.

The dissertation is based on my results published in the following papers:

1) Divya Shaji, Takayuki Amemiya, Ryotaro Koike, Motonori Ota, Interface property responsible for effective interactions of protean segments: Intrinsically disordered regions that undergo disorder-to-order transitions upon binding, *Biochemical and Biophysical Research Communications*, 478, 123-127 (2016).

2) Divya Shaji, The relationship between relative solvent accessible surface area (rASA) and irregular structures in protean segments (ProSs), *Bioinformation,* 12(9), 381-387 (2016)

Acknowledgements

I would like to express my very great appreciation to Prof. Motonori Ota for his valuable suggestions during the planning and development of this research work.

I am particularly grateful for the assistance given by Dr. Ryotaro Koike and Dr. Takayuki Amemiya.

I would like to thank the past and present lab members of the Ota Laboratory. They provided a pleasant environment that was a pleasure to work in.

I am grateful to the IDEAL development team, especially, Dr. Satoshi Fukuchi, for discussions.

Finally, I wish to thank my family for their support and encouragement throughout my study.

Abstract

Proteins that lack a well-defined conformation under native conditions are referred to as intrinsically disordered proteins. When interacting with partner proteins, short regions in disordered proteins can undergo disorder-to-order transitions upon binding; these regions are called protean segments (ProSs). It has been indicated that interactions of ProSs are effective: the number of contacts per residue of ProS interface is large. To reveal the properties of ProS interface that are responsible for the interaction efficiency, we classified the interface into core, rim and support, and analyzed them based on the relative accessible surface area (rASA). Despite the effective interactions, the ProS interface is mainly composed of rim residues, rather than core. The ProS rim is more effective than the rim of heterodimers, because the average rASAs of ProS rim, which is significantly large in the monomeric state, provides a large area to be used for the interactions. The amino acid composition of ProSs correlated well with those of heterodimers in both the core and rim. Therefore, the composition cannot explain why the rASAs of the ProS rim are large in the monomeric state. The balance between a small core and a large rim, and the large solvent exposure of the rim in the monomeric state, are the key to the disorder-to-order transition and the effective interactions of ProSs. The amino acid compositions and the relative accessible surface areas (rASAs) of ProS secondary structural elements (SSEs) at the interface, core and rim

were compared to those of heterodimers. The average number of contacts of alpha helices and irregular residues was calculated for each ProS and heterodimer. Furthermore, the ProSs were classified into high and low efficient based on their average number of contacts at the interface. The results indicate that the irregular structures of ProSs and heterodimers are significantly different. The rASA of irregular structures in the monomeric state (rASAm) is large, leads to the formation of larger ΔrASA and many contacts in ProSs.

Contents

List of Figures

List of Tables

Dedicated to my sweet son Abhinav Das

Chapter 1

Intrinsically Disordered Proteins (IDPs)

1.1 Introduction

Intrinsically Disordered Proteins (IDPs) or Intrinsically Unstructured Proteins (IUPs) are proteins that lack stable three dimensional structures under physiological conditions (Dunker et al., 2001; Dyson and Wright, 2005). These IDPs/IUPs are highly abundant in nature and have numerous biological activities (Uversky, 2014). IDPs are more abundant in eukaryotic proteomes than in archaea and prokaryotes (Ward et al., 2004).

Some proteins are predicted to be entirely disordered, while others contain disordered sequences, named to as intrinsically disordered regions (IDRs), in combination with structured globular domains. The majority of proteins in eukaryotic proteomes contain both intrinsically disordered and ordered regions (Wright and Dyson. 2015; Van Der Lee et al., 2014). There are two major classes of protein disorder; short regions and long regions. The short regions are typically <15-20 residues which serve as flexible linkers between or within domains, and long regions are >30–50 residues. These two classes have different amino acid propensities.

Proteins without intrinsically disordered regions (IDRs) are called structured or ordered proteins, and proteins with disordered sequences that do not adopt any tertiary structure are referred to as IDPs or IUPs (Van Der Lee et al.,2014).

1.2 Interactions of Disordered Proteins

Previous research showed that disordered segments adopt a largely extended and open conformation in the complex. Generally shorter regions undergo disorder-to-order transition than long regions. These short regions are usually below 100 residues; in many cases the length of the disordered binding regions are less than 30 residues. Compared to the globular proteins, the interface of disordered proteins is more hydrophobic, and the interaction contacts are also significantly different. IDPs tend to favor hydrophobic-hydrophobic contacts with the partner proteins at the interface (Mészáros et al., 2007).

Another research showed that polar and charged residues play a larger role in the interfaces of intrinsically disordered proteins compared to the interfaces of globular proteins. This suggests that polar interactions are key contributors to the specificity of interactions that involve intrinsically disordered proteins (Wong, Na, and Gsponer, 2013). Intrinsically disordered (ID) regions provide large surface area. By providing larger interaction area, ID regions can support interactions with several molecules to form large multimeric complexes (Gsponer and Babu, 2009).

1.3 Intrinsically Disordered Proteins in human diseases

Numerous IDPs are associated with human diseases, including cancer, cardiovascular diseases, amyloidoses, neurodegenerative diseases, and diabetes. IDPs, such as alpha-synuclein, tau protein, p53, and BRCA1, are attractive targets for drugs modulating protein-protein interactions (Uversky, Oldfield, and Dunker, 2008). Each of these diseases originates from the dysfunction of a particular protein. Some disease-related proteins have an intrinsic propensity to form pathologic conformation(s). For other proteins, interactions or impaired interactions with other proteins, small molecules, and other endogenous factors can induce conformational changes and increase the propensity to misfold. Misfolding and dysfunction can be caused by point mutation(s), impaired Post Translational Modifications (PTMs), an increased probability of degradation, impaired trafficking, loss of binding partners, or oxidative damage (Uversky, 2012; Uversky et al., 2014). A high percentage of cancer-associated proteins have long disordered regions (Iakoucheva et al., 2002).

1.4 General characteristics of Intrinsically Disordered Proteins(IDPs)

Intrinsically disordered proteins have low sequence complexity and amino acid compositional bias, with a low content of hydrophobic amino acids (Val, Leu, Ile, Met, Phe, Trp and Tyr), and a high content of particular polar and charged amino acids (Gln, Ser, Pro, Glu, Lys, Gly and Ala). The hydrophobic amino acids normally form the core of a folded globular protein (Dyson and Wright, 2005a; Romero et al., 2001; Vucetic et al., 2003).

1.5 Functions of Intrinsically Disordered Proteins

Intrinsically disordered proteins are frequently involved in key biological processes such as cell cycle control, membrane fusion and transport, transcriptional and translational regulation, and signal transduction (Iakoucheva et al., 2004; Wright and Dyson, 1999; Dyson and Wright, 2002; Dyson and Wright, 2005b; Minezaki et al., 2006). Intrinsically disordered regions are often involved in molecular recognition and protein modifications including phosphorylation (Iakoucheva et al., 2004; Dunker et al., 2002). Many intrinsically disordered proteins undergo transitions from disordered to ordered states on binding to their partners, known as coupled folding and binding mechanism (Dyson and Wright, 2005; Wright and Dyson, 1999; Dyson and Wright, 2002; Demchenko, 2001). Coupled folding and binding might involve just a few residues or an entire protein domain (Dyson and Wright, 2005; Zhou et al., 2001).

1.6 How to predict Disordered Proteins?

1.6.1 Experimental Methods

An experimental method for obtaining information about disordered proteins is NMR spectroscopy. Other techniques such as fluorescence spectroscopy, circular dichorism (CD), Raman spectroscopy and vibrational spectroscopy etc. can give important information about disordered proteins.

1.6.2 Computational methods

A number of computer programs are available for the prediction of disordered or unstructured regions from amino acid sequences. These include DisEMBL, GLOBPLOT, PONDR-FIT, FoldIndex and DISOPRED2 etc.

DisEMBL

DisEML is a computational tool for the prediction of disordered/unstructured regions within a protein sequence. This predictor uses three different criteria for assigning disorder: loops/coils, hot loops, i.e. coils with high temperature factors and missing coordinates in X-Ray structure (Linding et al., 2003).

GLOBPLOT

GlobPlot allows the user to plot the tendency within the query protein for order/globularity and disorder. It successfully identifies inter-domain segments containing linear motifs, and also apparently ordered regions that do not contain any recognized domain. The plots indicate that instances of known domains may often contain additional N- or C- terminal segments that appear ordered (Linding et al., 2003).

PONDR- FIT

PONDR- FIT is a meta-predictor of intrinsically disordered amino acids. This predictor introduced a consensus artificial neural network (ANN) prediction method, which was developed by combining the outputs of several individual disorder predictors. PONDR-FIT, was found to improve the prediction accuracy over a range of 3 to 20% with an average of 11%

compared to the single predictors, depending on the datasets being used (Xue et al., 2010).

FoldIndex

FoldIndex is a graphic web server, predicts a given protein sequence is intrinsically unfolded or not, which is based on the average residue hydrophobicity and net charge of the sequence (Prilusky et al.,2005).

DISOPRED

The DISOPRED server allows users to submit a protein sequence, and returns a probability estimate of each residue in the sequence being disordered. The results are sent in both plain text and graphical formats, and the server can also supply predictions of secondary structure to provide further structural information (Ward et al., 2004).

1.7 Databases of Disordered Proteins

A number of databases of protein disorder have recently been established. Some of them are shown below.

1.7.1 Disprot

The Database of Protein Disorder (DisProt) is a curated database that provides information about proteins that lack fixed 3D structure in their putatively native states, either in their entirety or in part. The database includes the location of the experimentally determined disordered regions in a protein and also shows the methods used for disorder characterization (Sickmeier et al., 2007).

1.7.2 D2P2

The Database of Disordered Protein Prediction (D2P2) database aims to provide unified and exhaustive disorder predictions for all currently sequenced genomes with protein annotations. The disorder/structure annotations of this database enable comparison of the disorder predictors with each other and examination of the overlap between disordered predictions and SCOP domains on a large scale (Oates et al., 2013).

1.7.3 IDEAL

IDEAL provides the experimentally verified intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs). IDEAL contains manually curated annotations on IDPs in locations, structures, and functional sites such as protein binding regions and post translational modification sites together with references and structural domain assignments (Fukuchi et al., 2012; Fukuchi et al., 2014).

1.7.4 PED

PED is an openly accessible database for the deposition of structural information on IDP and denatured protein ensembles based on Nuclear Magnetic Resonance (NMR) and Small-angle X-ray Scattering (SAXS) data. The deposition of structural coordinates as well as primary data can be used for evaluating and re-calculating the ensembles (Varadi et al., 2014).

1.7.5 MobiDB

MobiDB was designed to offer a centralized resource for annotations of intrinsic protein disorder. This database features three levels of annotation: manually curated, indirect and predicted. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest (Di Domenico et al., 2012).

The DisProt, IDEAL, MobiDB, and PED databases collect experimentally verified disordered regions and proteins.

1.8 Protean Segments (ProSs)

Protean segments are short regions in disordered proteins that can undergo disorder to order transitions upon binding to their partners. This phenomenon is known as coupled folding and binding in which a short flexible segment binds to its binding partner by forming a specific structure which acts as the molecular recognition element. Although ProS, MoRF (Molecular Recognition Features) and ELM (Eukaryotic Linear Motifs) are similar concepts, MoRF has a length limitation of 70 residues and an ELM should have a motif that can be described in a regular expression (Fukuchi et al., 2012; Fukuchi et al.,2014).

The definition of ProS depends only on evidence of a disorder-to-order transition. ProS can include IDRs whose structures are induced upon binding to small ligands. ProSs do not necessarily assume secondary structures in the binding state, and long IDRs or IDRs without a motif can also be ProSs. Long intrinsically disordered regions (IDRs), such as p27Kip1 (PDB: 1jsu) and Tcf3 (PDB: 1g3j), can transform into ordered states (Tompa et al., 2009). ProSs can also cover these IDRs (Fukuchi et al., 2012; Fukuchi et al., 2014).

1.9 Aim

The aim of this work is to investigate the interaction efficiency of Protean segments (ProSs) to those of ordered proteins (heterodimers). For that, the amino acid residues in ProSs and heterodimers were classified into surface, interior and interface based on their relative solvent accessible surface area (rASA). The interfaces of ProSs and heterodimers were further classified into core, rim and support. To examine the efficiency of interactions, the average number of contacts of ProSs at the interface, core and rim was calculated and compared to those of heterodimers.

Chapter 2

Analysis of the interface residues in Protean Segments (ProSs)

2.1 Introduction

Protein–protein interactions play an important role in many biological functions. To study the interaction between two proteins, a crystal structure of the protein–protein complex is necessary. The protein structures have been solved by NMR and X-ray crystallography and have been deposited into the Protein Data Bank (PDB). The characteristics of the protein-protein interfaces can be calculated from the atomic coordinates of the protein-protein complex available from the Protein Data Bank (PDB). (Chothia and Janin, 1975; Reich mann et al., 2007; Yan et al., 2008).

An interface is the interacting region of protein–protein complexes. The interacting residues become buried during the complex formation. A large interface is needed for strong interactions. Larger interfaces are usually formed between permanent complexes (Jones and Thornton, 1996; Mészáros et al., 2007). Many researches have been confirmed that, hydrophobic interactions play an important role in protein-protein interactions (Young, Jernigan, and Covell, 1994; Berchanski, Shapira, and Eisenstein, 2004; Yan et al., 2008). Another important property of interfaces is the existence of "hot-spot" residues, that make the largest contributions to complex formation (Keskin, Ma, and Nussinov, 2005).

IDPs usually use short segments of IDRs that undergo disorder-to-order transitions upon binding to their partners (i.e., coupled folding and binding) (Dyson and Wright, 2002; Wright and Dyson, 1999). We call these short segments protean segments (ProSs) (Fukuchi et al., 2012; Fukuchi et al., 2014).

The concepts of molecular recognition features (MoRFs) (Mohan et al., 2006; Vacic et al., 2007; Oldfield et al., 2005) and eukaryotic linear motifs (ELMs) or short linear motifs (SLiMs) (Dinkel et al., 2013; Davey et al., 2012) are similar to ProSs, but the definitions are partially different from each other (Fukuchi et al., 2012; Fukuchi et al., 2014).

As such binding regions (e.g., ProSs) are essential for the molecular function of IDPs, more attention has been paid to their interactions, and several characteristics have been revealed (Mohan et al., 2006; Vacic et al., 2007; Oldfield et al., 2005; Cheng et al., 2007; Mészáros et al., 2007; Wong, Na, and Gsponer, 2013; Fuxreiter et al., 2004; Galea et al., 2008; Hsu et al., 2013). In particular, it has been indicated that the interactions of ProSs are effective (Mészáros et al., 2007): on average, the number of contacts of ProS interface with its interaction partners is larger than that of globular proteins (e.g., heterodimers).

This has been explained by their unique interaction mode employing coupled folding and binding (Mészáros et al., 2007), but the details are still unclear. In this study, we focused on the interface of ProSs and compared it with that of heterodimers. We have investigated separately the characteristics of ProS interfaces and heterodimer interfaces. The interface residues were further classified into core, rim and support (Wong, Na, and Gsponer, 2013; Levy, 2010) and their relative solvent accessible surface areas (rASA) were analyzed in detail.

The residues in the interface core are the most buried residues upon protein binding, generally at the central region of the interface, and play an important role in the interaction (Levy, 2010), like hot spots (Keskin, Ma, and Nussinov, 2005; Guharoy and Chakrabarti, 2005). The residues in the interface rim are located on the outer edges of the interface that remain partially exposed to the solvent (Levy, 2010). The support residues play an insignificant role in the interaction, which represents the intersection between the interior and the interface.

Comparisons show significant differences between the two types of interfaces in interface size, rASA and average number of contacts. The major finding of our work is that the ProS interface is mainly composed of rim residues even though it has effective interactions. The key to effective interactions of ProSs is the solvent exposure of rim residues in the monomeric state.

2.2 Materials and Methods

2.2.1 ProSs and heterodimers

All ProSs (210) in 70 protein sequences were collected from the IDEAL database (as of August 2013) (Fukuchi et al., 2012; Fukuchi et al., 2014).If more than one ProS were found in a protein and their positions overlapped, we chose the longest ProS. The sequence redundancy was removed with 80% sequence similarity based on the CLUSTALW alignment (Thompson, Higgins, and Gibson, 1994). Hierarchical clustering was done with R (Ihaka and

Gentleman, 1996) using complete-linkage clustering and the longest ProS in a cluster was selected as the representatives. A non-redundant set contained 99 ProSs. DNA-binding ProSs and one-to-many binding ProSs (a single ProS binds to two or more different partners, (Hsu et al., 2013)) were discarded. Both the X-ray and NMR structures were used in this study.

A non-redundant dataset of 276 heterodimers was selected from the Protein DataBank (PDB) (Berman et al., 2000), using the PDB's advanced search interface (as of July 2014).

The search criteria satisfied the following conditions: (1) less than 30% sequence identity; (2) the macromolecule type contained only proteins; (3) the oligomeric state was heterodimer; (4) each chain was greater than 100 residues; and (5) structures determined by X-ray crystallography had higher than 3 Å resolutions. Only smaller protomers were analyzed as the reference of ProSs.

2.2.2 Amino acid propensity

The propensities of amino acids are represented as the Chou–Fasman parameters (Chou and Fasman, 1978), $CF(a,P) = \frac{N^{a}(P)/N(P)}{N_{all}^{a}/N_{all}}$, where N^{a} *(P)* is the number of amino acid residue *a* in place *P*, *N (P)* is the total number of residues in *P*, N_{all}^{a} is the total number of amino acid residue *a* in the protein sequence, and N_{all} is the total number of residues in the protein sequence. In *P*, we considered the interface, core and rim residues in ProSs and heterodimers.

To calculate the reference states (the denominator), we used SCOP25 proteins (version 1.75) (Murzin et al., 1995).

2.2.3 Relative ASA and residue contact

We classified the residues into surface, interior and interface. Based on the definitions by Levy (Levy, 2010), the interfaces were further classified into core, rim and support. The relative solvent accessible surface area (rASA) is defined as the total accessible surface area (ASA) of the residues in a protein structure normalized by the ASA of the residues in the most exposed state to a solvent molecule, generally water (Rose et al.,1985).

The rASAs of each residue, in the monomeric and complex states (rASAm and rASAc, respectively) were computed for ProSs and heterodimers using the program Naccess (Hubbard and Thornton, 1993), which is an implementation of Lee and Richard's algorithm (Lee and Richards, 1971). ΔrASA is the difference between the rASAs of monomeric and complex states. The rASAs were averaged for interface, core and rim residues, to derive the average rASAs of proteins.

Two residues, i and j, were considered to be in contact if any atom of residue i was within a distance of < 4.5 Å with any atom of residue j (Heringa and Argos, 1991; Nath Jha, Vishveshwara, and Banavar, 2010). We calculated the number of external contacts for ProSs and heterodimers at the interface, core and rim. External contacts are defined as the contacts between the proteins and their interaction partners. The average number of contacts at the interface, core and rim was calculated for each ProS and heterodimer.

2.2.4 Statistical analysis

Wilcoxon rank-sum test was performed by RStudio (Racine, 2012) to calculate the P-values (Table 2.1). $P < 0.01$ was considered statistically significant.

2.3 Results and Discussion

2.3.1 Composition of interfaces and effective interactions of ProSs

Based on the protein dimeric structures, amino acid residues in ProSs and heterodimers were classified into surface, interior and interface residues (Levy, 2010) (Figure 2.1). Surface residues are the exposed residues and interior residues are the buried residues. The residues in the interior are more hydrophobic than the residues on the surface. An interface is the interacting region of protein–protein complexes and represents the intersection between surface and interior (Jones and Thornton, 1996; Mészáros et al., 2007).

As in nature ProSs have a small number of intra-chain contacts, and only adopt structures when interacting with partner proteins, ProSs have a larger number of interface residues and a smaller number of interior residues than heterodimers.

Figures 2.2A and B further break down the composition of the interface residues into core, rim and support residues in ProS and heterodimer interfaces, respectively.

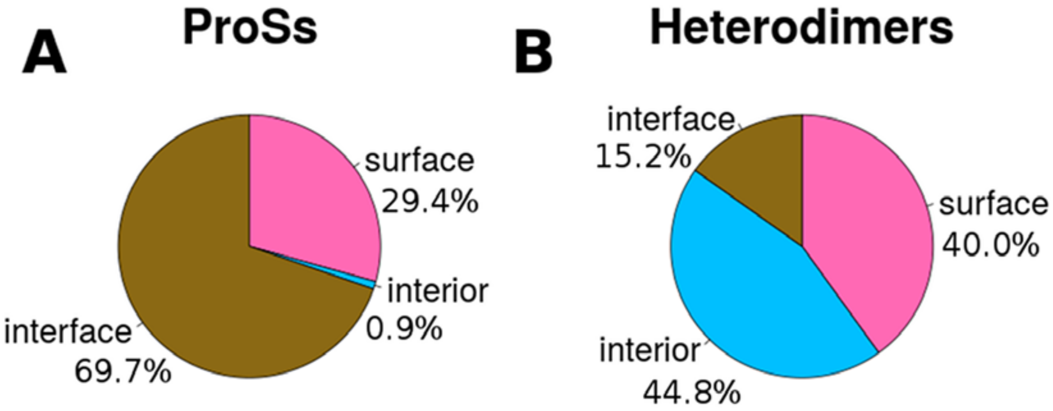

FIGURE 2.1: Composition of the residues in ProSs and heterodimers. The composition of the surface (pink), interior (blue) and interface (goldenrod) residues in ProSs (A) and in heterodimers (B).

In the ProS interface, core residues are less abundant (33.7%) compared with the heterodimer interface (36.8%). Moreover, in ProSs, the interface is mainly composed of rim (64.7%), which is nearly double that observed in heterodimers (35.3%). The distribution of the rates of core and rim (Fig. 2.3) is significantly different in ProSs and heterodimers as assessed by the Wilcoxon rank-sum test (P-values: core = 4.5e-05, rim =1.3e-40).

In summary, the ProS interface is composed of a small core and a large rim. This statement based on the rates gives a slightly different representation from the results of a previous report on absolute values (Wong, Na, and Gsponer, 2013), denoting that the number of residues in the core of the 1D segment (corresponding to ProS) is smaller than that of the 3D complex proteins (heterodimers), but in the rim, the numbers of residues are almost equal.

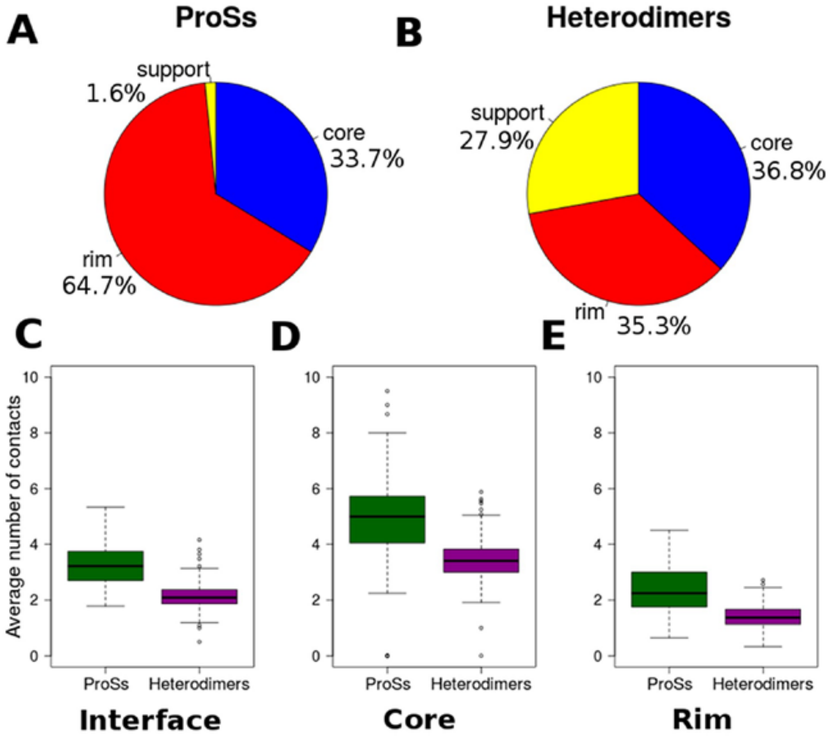

FIGURE 2.2: Composition and interaction efficiency of interface residues (defined by ΔrASA > 0). Composition of interface residues in ProSs (A) and heterodimers (B). Core, rim and support are shown in blue, red and yellow, respectively. Box-plots of the average number of contacts of the ProSs and heterodimers at the interface (C), core (D) and rim (E). The distributions of ProSs and heterodimers are colored in green and purple, respectively. The differences between the distributions were evaluated, and the P-values are shown in Table 2.1. The residues in the interface core (rASAm > 25% and rASAc < 25%) are the most buried residues upon protein binding and are generally at the central region of the interface. The residues on the outer edges of the interface that remain partially exposed to solvent are a part of the interface rim [rASAc (and rASAm)>25%]. Support (rASAm<25%) represents the intersection between the interior and the interface (Levy, 2010).

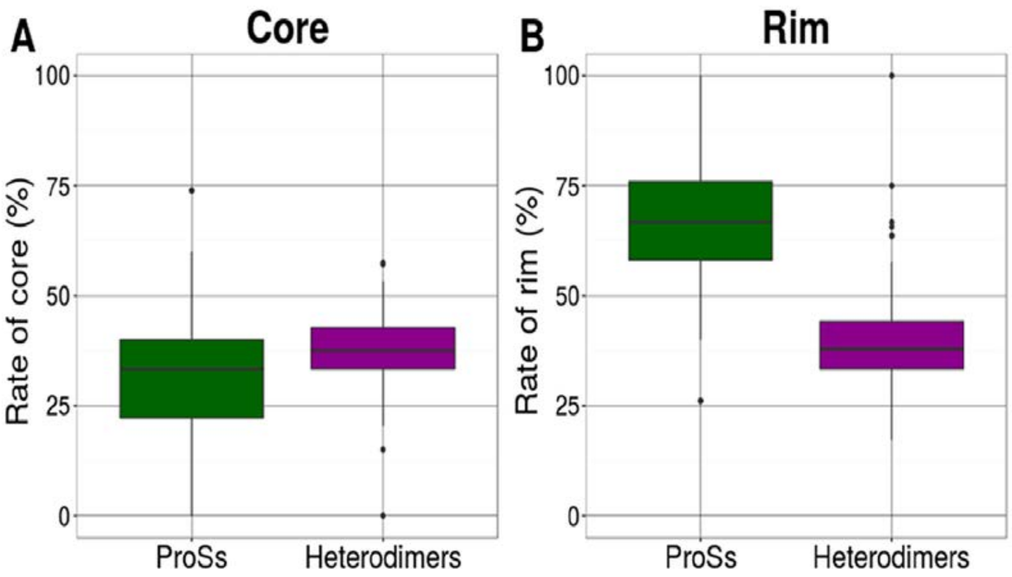

FIGURE 2.3: Box-plots of the rates of the core and rim residues at the interface of ProSs and heterodimers. (A) Distribution of the core residues in ProSs (green) and heterodimers (purple). (B) Distribution of the rim residues in ProSs and heterodimers. The distributions are significantly different as assessed by the Wilcoxon rank-sum test (core = 4.50e-05, rim = 1.28e-40).

To examine the efficiency of interactions, we calculated the average number of (inter-chain) contacts of interface residues for each ProS and heterodimer, and compared their distributions. As was shown in Fig. 2.2C as well as in a previous report (Mészáros et al., 2007), the ProS interface can be in contact with a larger number of residues of the interaction partners compared with the heterodimer interface, confirming that the ProS interaction is effective. However, this result seems to be inconsistent with our results of the ProS interface composition (Fig. 2.2A and B), because for effective interactions, a large core and a small rim are expected. To analyze the contribution of residues, the average number of contacts was derived individually for the core and the rim of the interfaces (Fig. 2.2D and E). Apparently, on average core residues have a larger number of contacts compared with the rim, confirming that having a core should be reasonable for effective interactions. Moreover, it is noticeable that in both the core and rim cases, the average number of contacts by ProS residues is larger than that by heterodimers (see the P-values in Table 2.1). This indicates that ProS residues contribute to effective interactions not only through the core, but also through the rim. In particular, because of their abundance, the efficiency of the ProS rim is remarkable.

To prove this hypothesis, we ignored the interactions of the core, rim or support individually, and calculated the average number of contacts again (Fig. 2.4). When we took into account the rim contacts and ignored those of the core and support, the average number of contacts of the ProS was different from that of the heterodimer (Fig. 2.4A and C). Only when we ignored the rim contacts, the average number of ProS contacts was almost equal to that of the heterodimer (Fig. 2.4B), indicating that the contribution of the rim to interactions is significant, and the interaction mechanism of the region should be addressed.

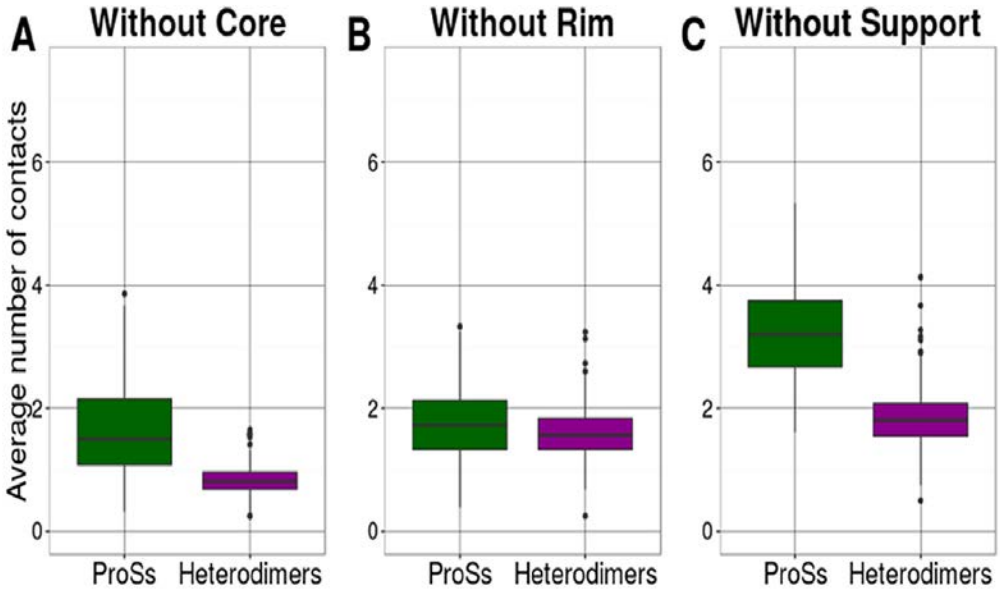

FIGURE 2.4: Box-plots of the average number of contacts of ProSs and heterodimers ignoring the interactions of core, rim or support. (A) Average number of contacts without core. (B) Average number of contacts without rim. (C) Average number of contacts without support. ProSs and heterodimers are colored in green and purple, respectively. The distributions were evaluated by the Wilcoxon rank-sum test. The P-values are 4.42e-23, 0.02 and 2.52e-39 for A, B and C, respectively.

2.3.2 Relative ASA analyses

It is well known that the number of contacts between proteins, the interaction energy of proteins, and the interface area of proteins are almost proportional in the protein-protein interactions (Ooi et al., 1987; Eisenberg and McLachlan, 1985).This means that ΔASA, defined by the difference between ASAs in the unbound (monomeric) and in the bound (complex) states, is a good indicator of the degree of interactions or the number of contacts.

The relationship is unchanged if the values are normalized by the number of interface residues. In fact in our data, the average number of contacts and the average ΔASA showed a good correlation (Fig. 2.5A).

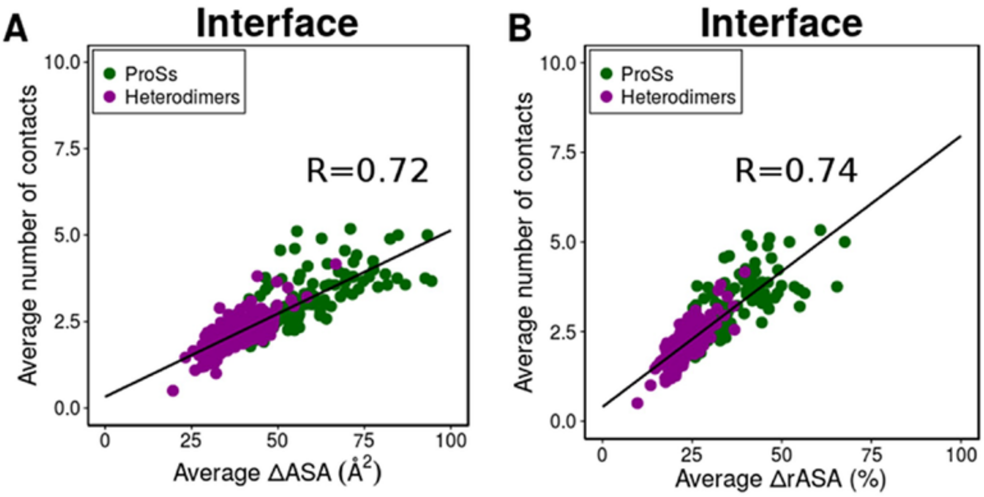

FIGURE 2.5: **Scatterplots of the average ΔASA and average ΔrASA vs. the average number of contacts.** (A) Average ΔASA vs. the average number of contacts of ProSs (green) and heterodimers (purple) at the interface. (B) Average ΔrASA vs. the average number of contacts of ProSs and heterodimers at the interface. The ΔrASA upon protein binding was calculated by subtracting the rASA of monomers (monomeric state) and that of their complexes (complexed state) (Levy, 2010). The average number of contacts of each ProS and heterodimer was calculated.

We confirmed that when we used ΔrASA instead of ΔASA the same relationship was held (Fig. 2.5B). ΔrASA of each residue is defined by the difference between rASA of the monomeric state (rASAm) and that of the complex state (rASAc), and both rASAs are used to define the core, rim and support residues. Therefore, analyzing rASA is promising for connecting the feature of the ProS rim with its effective interactions.

As shown in Figure 2.5B, ΔrASAs were averaged over the interface residues, correlates well with the average number of contacts. This indicates that when the average number of contacts of ProSs is larger than that of heterodimers (Fig. 2.2C–E), it would be caused by a larger average rASAm of ProS, a smaller average rASAc of ProS, or both.

In Fig. 2.6A–C and 2.6D–F, the distribution of the average rASAm of ProSs and average rASAc of ProSs is shown, respectively, for the interface, core and rim, and compared with those of heterodimers. In both the core and rim, ProSs in the monomeric state are more exposed to water than heterodimers (Fig. 2.6B and C), resulting in more solvent exposure of the ProS interfaces (Fig. 2.6A).

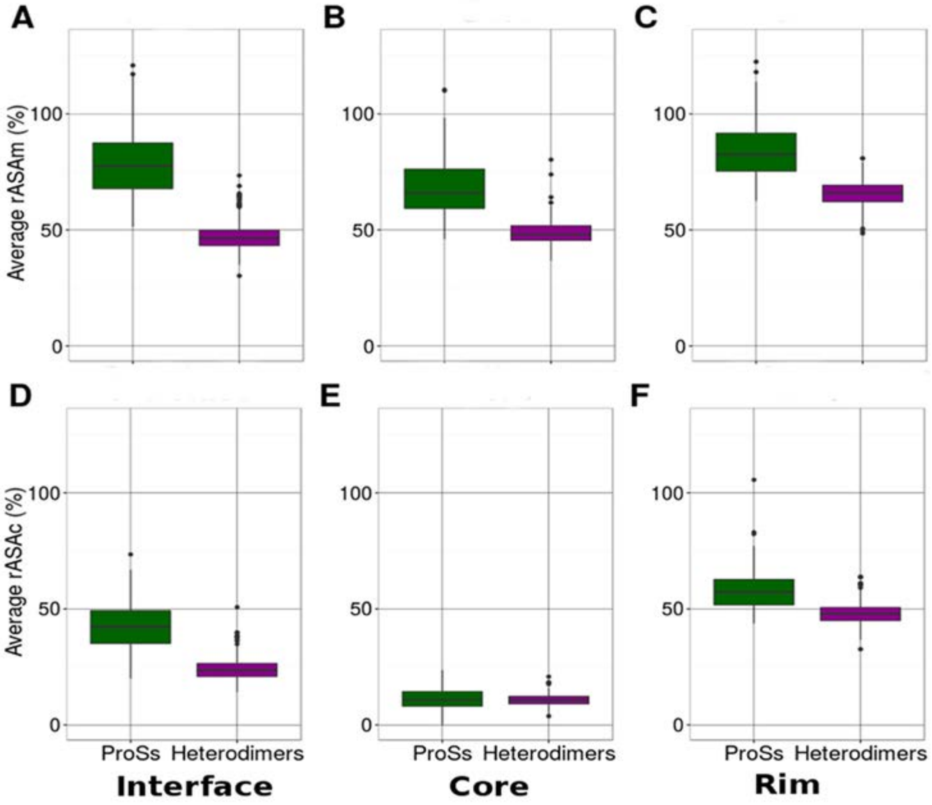

FIGURE 2.6: Average rASAs of ProSs and heterodimers. Average rASAm at the interface (A), core (B) and rim (C). Average rASAc at the interface (D), core (E) and rim (F). The differences between the distributions were evaluated, and the P-values are shown in Table 2.1.

The differences are confirmed by a statistical test (Table 2.1). In contrast, in the complex state, the solvent exposure of the core of the ProSs and heterodimers is small and has similar values (Fig. 2.6E), reflecting the definition of Levy (Levy, 2010): the core residues should be buried in the complex state (rASAc < 25%). In contrast, in the rim the rASAc values of ProSs are larger than those of heterodimers (Fig. 2.6F). Although a larger rASAc is disadvantageous for effective interaction, the solvent exposure of the ProS rim in the monomeric state (rASAm) is fairly large (see Fig. 2.6C and P-values in Table 2.1), resulting in a larger ΔrASA, or numerous contacts. Contour plots of average rASAm and rASAc are shown in Fig. 2.7.

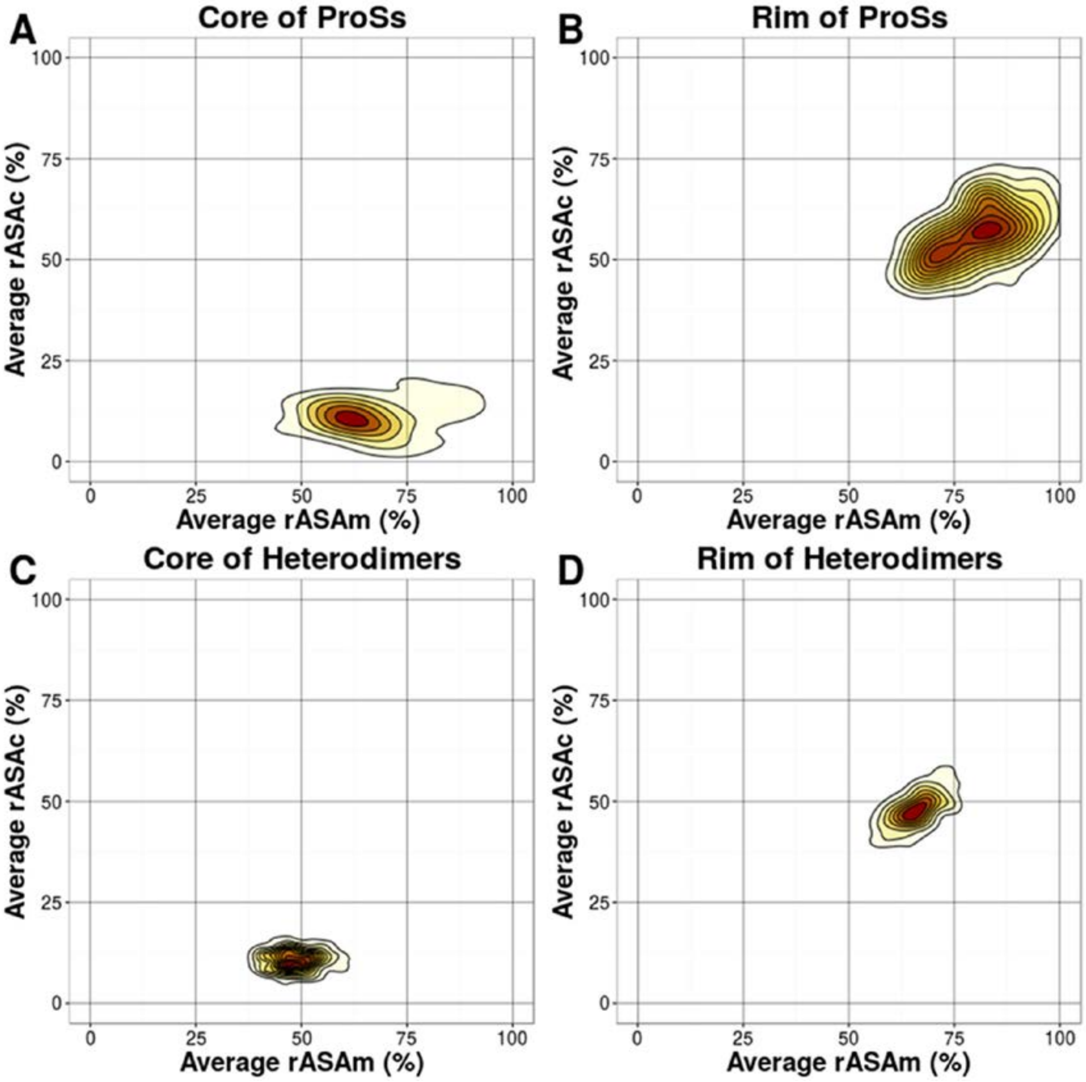

FIGURE 2.7: Contour plots of the average rASAm and average rASAc. (A) Average rASAm vs. average rASAc of the ProS core. (B) Average rASAm vs. average rASAc of the ProS rim. (C) Average rASAm vs. average rASAc of the heterodimer core. (D) Average rASAm vs. average rASAc of the heterodimer rim. The rASAs of each residue in the monomeric and in the complexed states in ProSs and heterodimers were calculated using Naccess (Hubbard and Thornton, 1993). The highest density regions are shown in red, and the lowest density regions are in white. The distance from the diagonal line represents the ΔrASA.

2.3.3 Amino acid compositions of the core and rim

As structural features are frequently explained by the amino acid composition, the amino acid compositions of ProSs vs. heterodimers were examined. We computed the Chou–Fasman parameters (Chou and Fasman, 1978) for core and rim residues. In Fig. 2.8, the correlations between ProS core vs. heterodimer core, and ProS rim vs. heterodimer rim are indicated.

In both cases, positive correlations were observed with 0.41 and 0.78 correlation coefficients for core and rim residues, respectively. This indicates that the amino acid composition of the ProSs core/rim is similar to that of heterodimers. Especially, the rim composition of ProSs and heterodimers is quite similar (Fig. 2.8B), suggesting that the amino acid composition is not the determinant of the large solvent exposure of the ProS rim in the unbound state.

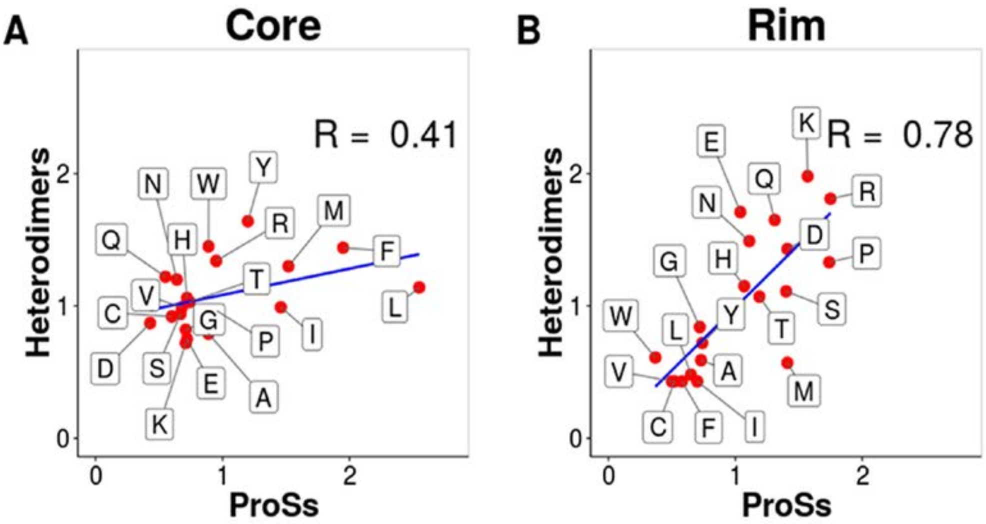

FIGURE 2.8: Scatter plots of the Chau–Fasman parameters (Chou and Fasman, 1978). (A) ProS core vs. heterodimer core. (B) ProS rim vs. heterodimer rim.

To characterize the amino acid compositions of the core and rim, we computed the Chou–Fasman parameters (Chou and Fasman, 1978) of the protein interior (Levy, 2010) and IDR using the monomeric proteins in Protein Quaternary Structure Fileserver (PQS) (Henrick and Thornton, 1998) (sequence identity < 25%), and compared them with those of the core and rim (Figs. 2.9 and 2.10). The core and rim residues correlated well with the protein interior and IDR, respectively, implying that the core residues are essentially hydrophobic and the rim residues are polar. This result is consistent with that of Wong et al. (Wong, Na, and Gsponer, 2013), in which the significance of electrostatic interactions in the rim was emphasized.

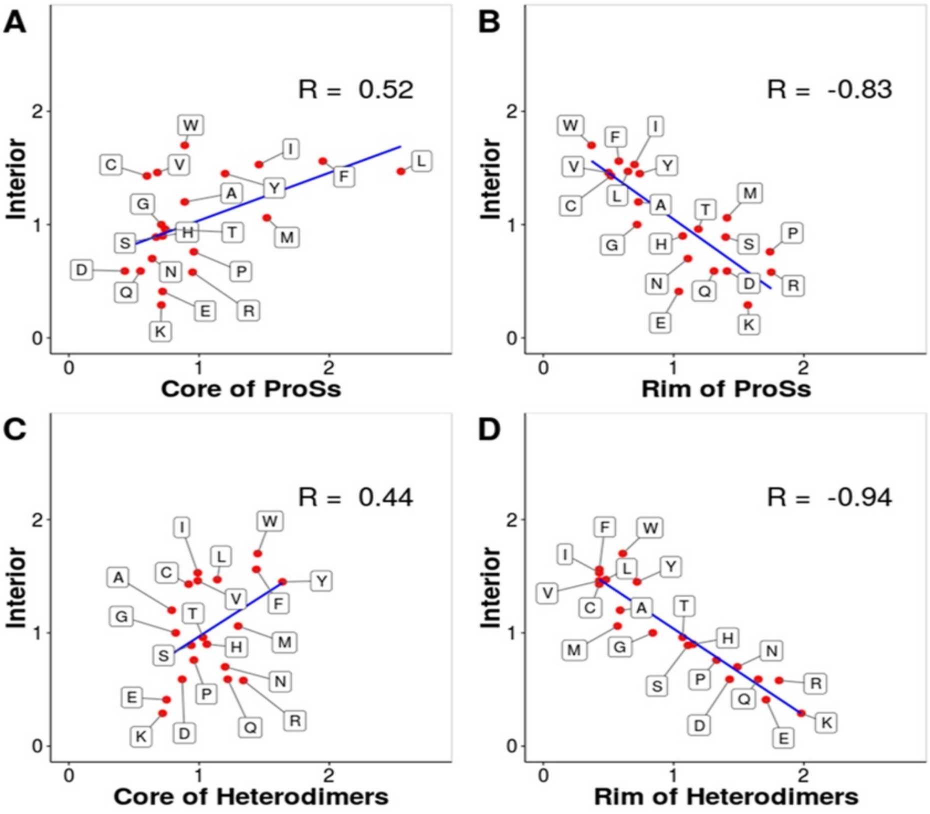

FIGURE 2.9: Scatter plots of the core and rim vs. the interior. (A) ProS core vs. interior. (B) ProS rim vs. interior. (C) Heterodimer core vs. interior. (D) Heterodimer rim vs. interior. The amino acid propensities were calculated using the Chou–Fasman formula (Chou and Fasman, 1978). The monomeric proteins were selected from PQS (Henrick and Thornton, 1998) with 25% sequence identity. The monomeric residues were classified into surface and interior based on their rASAm with a 25% cutoff (Levy, 2010)

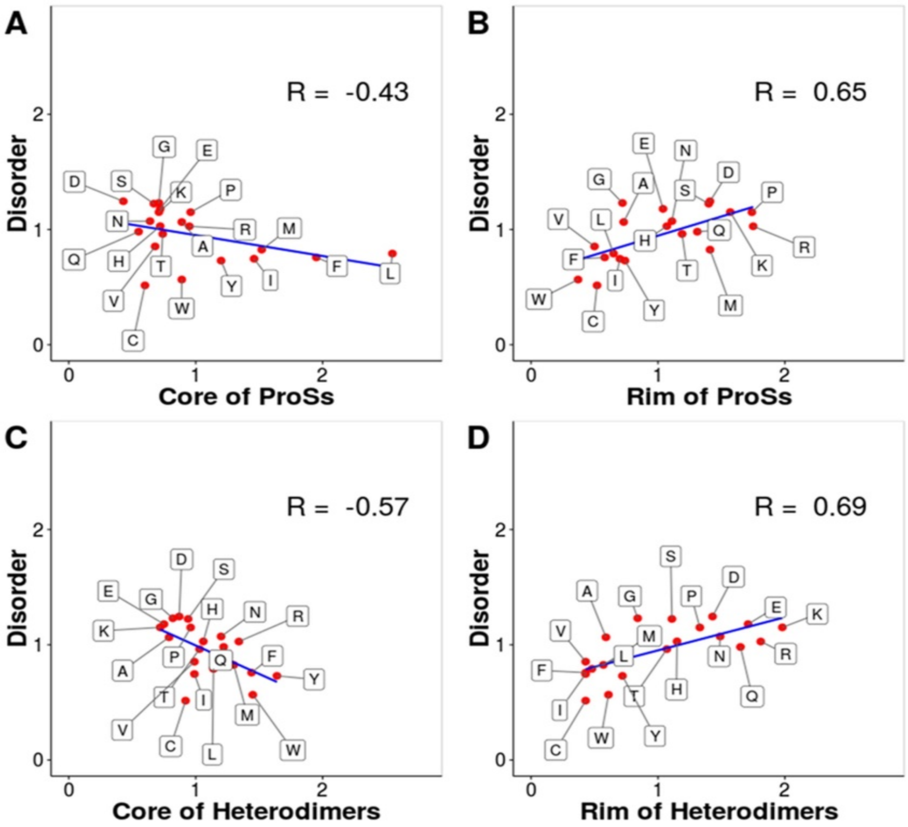

FIGURE 2.10: Scatter plots of the core and rim vs. the disordered residues. (A) ProS core vs. disordered residues. (B) ProS rim vs. disorder. (C) Heterodimer core vs. disordered residues. (D) Heterodimer rim vs. disorder. The amino acid propensities were calculated using the Chou–Fasman formula (Chou and Fasman, 1978). The disordered residues were extracted from SCOP25 based on the SEQRES and SEQATM annotations (Murzin et al., 1995).

Considering that ProSs are disordered in the unbound state and the ProS interface is mainly composed of rim, the similarity between the rim and IDR is reasonable. In other words, in addition to the ProS surface, the ProS rim is a significant area to ProS being disordered in the unbound state. Furthermore, the ProS rim contributes to effective interactions in the bound state using a large surface area in the unbound state. The relevant balance between a small core and a large rim in the ProS interface is crucial for modulating the disorder-to-order transition appropriately.

Table 2.1: ***P*-values of ProSs and heterodimers (using the Wilcoxon rank- sum test)**

Features	Places	P-values
Average number of contacts	Interface	1.99e-32
	Core	4.76e-25
	Rim	8.93e-22
Average rASAm	Interface	5.60e-48
	Core	2.25e-37
	Rim	4.59e-37
Average rASAc	Interface	3.06e-42
	Core	0.35
	Rim	7.51e-27

2.4 Conclusion

We investigated the characteristics of the ProS interface and compared them with those of heterodimers. Our analyses revealed that (a) the ProS interface is mainly composed of rim residues; (b) ProSs have a larger number of average contacts than heterodimers, and the contribution of the rim to effective interactions is significant; (c) the ProS rim has a larger average rASA in the monomeric state, which is the reason for why the ProS rim can interact with a larger number of residues in partner proteins, and (d) the amino acid composition in the core and rim is mostly unchanged in ProSs and heterodimers, and the cores are hydrophobic and the rims are polar.

In summary, the ProS rim is essential for effective interactions of ProSs via both its dominance in the ProS interface and its capability for more contacts using the large solvent exposure in the monomeric state. Furthermore, the small core and the large rim in the ProS interface, i.e., the well-balanced mixture of core and rim, is the key to the disorder-to-order transitions.

Chapter 3

Analysis of the Secondary Structure Elements (SSEs) in ProSs

3.1 Introduction

The goal of this work is to investigate the properties of secondary structure elements (SSEs) at the interface of ProSs relative to those of heterodimers. The interfaces of ProSs and heterodimers were classified into the core, rim, and support based on their relative solvent accessible surface area (rASA) (Levy, 2010). The average number of contacts of alpha helices and irregular residues was calculated for each ProS and heterodimer. Furthermore, the ProSs were classified into high and low efficient ProSs based on their average number of contacts at the interface. Compared to heterodimers, irregular residues of ProSs have larger number of contacts than their alpha helices. Moreover, irregular residues of ProSs have larger ΔrASA than their alpha helices. The rASA of irregular structures in the monomeric state is large, that leads to the formation of larger ΔrASA and many contacts in ProSs. In addition, high efficient ProSs have larger average rASA in the monomeric state (rASAm) and larger average ΔrASA, than low efficient ProSs.

3.2 Materials and Methods

3.2.1 ProSs and heterodimers

A non-redundant dataset of 99 ProSs and 276 heterodimers was used in this study. The creation of the datasets was described in the methods section of chapter2.

3.2.2 Secondary structure analysis

The program DSSP (Kabsch and Sander, 1983) was used to assign secondary structures. The eight types calculated by DSSP were reduced to three, such as alpha helices (H, G and I), beta strands (E) and irregulars (B, S, T and C). The amino acid propensity, average number of contacts and relative solvent accessible surface areas (rASAs) of alpha helices and irregulars were analyzed in detail.

3.2.3 Calculation of amino acid propensities

The amino acid propensities of the alpha helix and irregular residues were calculated using Chou-Fasman parameters (Chou and Fasman, 1978) (See Methods in Chapter2). To calculate the reference states (the denominator), the same secondary structure types of PDBSelect25 (Hobohm et al., 1992) proteins were used. PDBSelect25 contains a representative set of PDB entries with less than 25% sequence identity.

3.2.4 High and low efficient ProSs

Based on the average number of contacts in the interface, the ProSs were classified into high and low efficient ProSs. High and low efficient ProSs were defined as the contacts of ProSs with greater than 4 and less than 2.5, respectively. Short ProSs (less than 11 residues) were discarded from this classification. Several properties were analyzed for each high and low efficient ProSs (See Results and Discussion). The datasets contain 11 and 14 ProSs for high and low efficient, respectively. The radius of gyration (Rg) was calculated using Bio3D package (Grant et al., 2006) in R (Ihaka and Gentleman,1996).

3.3 Results and Discussion

3.3.1 Secondary structure analysis of ProSs and heterodimers

The secondary structure assignments for each of the ProS and heterodimer interface were determined by the DSSP program (Kabsch and Sander, 1983). This analysis (See Figs. 3.1A and B) showed that 33% of the residues in the ProSs dataset were alpha helices, 6% were beta strands, and 61% were residues of the irregular structure. The secondary structure distribution of ProSs interface is very different from those of heterodimers.

The content of irregular structures and beta strands are the largest difference between ProSs and heterodimers. Alpha helices are almost equally abundant in both data sets. ProS interface contains 15% more irregular residues, 13% fewer beta strands and 2% fewer alpha helices than heterodimers.

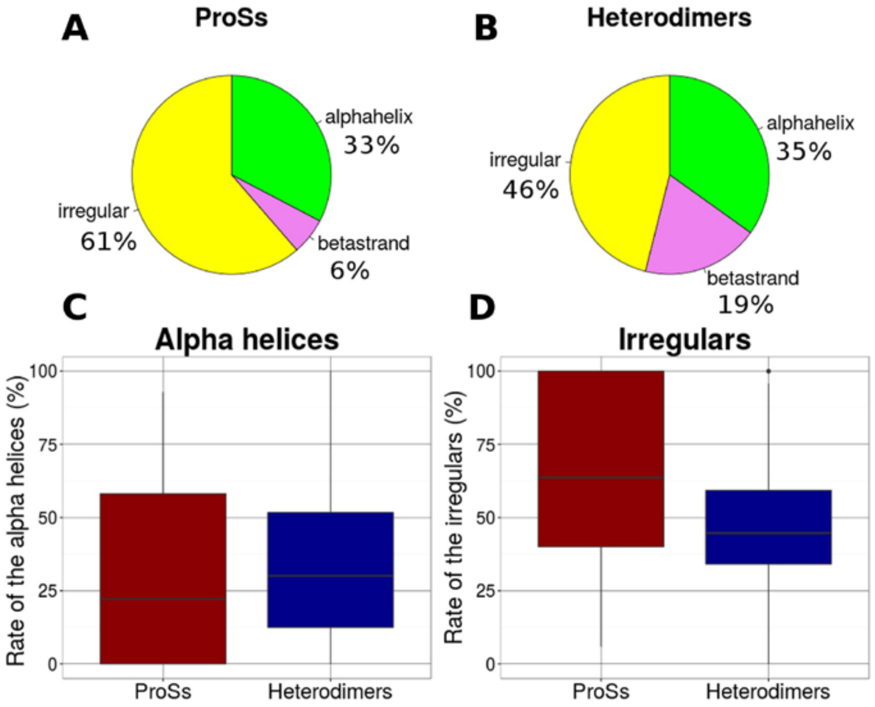

FIGURE 3.1: Distribution of secondary structure elements (SSEs) in ProS and heterodimer interface. The composition of secondary structure elements (SSEs) in ProS interface (A) and heterodimer interface (B). The program DSSP was used to assign secondary structures. The eight types calculated by DSSP (Kabsch and Sander, 1983) were reduced to three, such as alpha helices, beta strands, and irregulars. The distributions of alpha helices, beta strands and irregulars are colored in green, violet and yellow, respectively. Because of the shortage of beta strand residues in ProSs, alpha helices and irregulars were considered for further analysis. Box-plots of the rates of (C) alpha helix residues in ProSs (red) and heterodimers (blue) interface (D) irregular residues in ProSs and heterodimers interface. The distribution of the irregulars is significantly different as assessed by the Wilcoxon rank-sum test (alpha helices = 0.03, irregulars =1.05e-07).

The differences between the distributions were evaluated, and the boxplots of the rates of alpha helices and irregulars are shown in Fig. 3.1C and D. The alpha helix residues of ProSs and heterodimers are not significantly different (P-value = 0.03). It is important to note that, the irregular structures of ProSs and heterodimers are significantly different (P-value =1.05e-07).

3.3.2 Interactions of secondary structure elements (SSEs)

The amino acid propensities of the different secondary structure elements (SSEs) (alpha helices and irregular structures) for ProSs vs. heterodimers were examined. The Chou–Fasman parameters (Chou and Fasman, 1978) for alpha helix and irregular residues at the interface were calculated.

In Figs. 3.2A and B, the correlations between ProS alpha helices vs. heterodimer alpha helices and ProS irregulars vs. heterodimer irregulars at the interface are indicated. In both cases, positive correlations were observed with 0.50 and 0.61 for alpha helix and irregular residues, respectively. This indicates that the amino acid composition of the ProSs secondary structural elements (SSEs) is moderately similar to that of heterodimers.

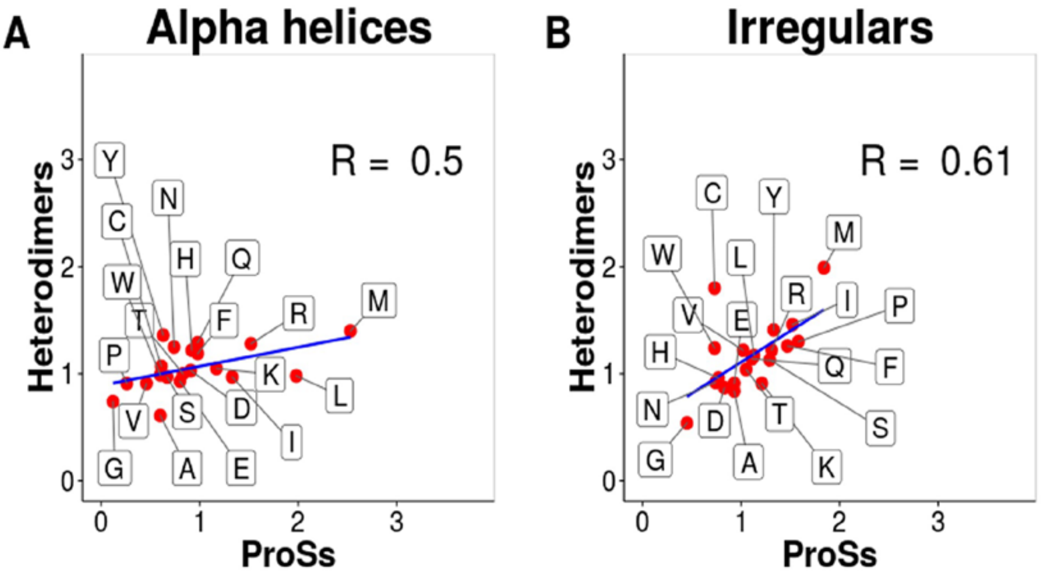

FIGURE 3.2: Scatter plots of the Chau–Fasman parameters (Chou and Fasman, 1978) of alpha helices and irregulars at the interface (A) ProS alpha helices vs. heterodimer alpha helices. (B) ProS irregulars vs. heterodimer irregulars.

Previous studies have indicated that the ProS interface can be in contact with a larger number of residues of the interaction partners compared with the heterodimer interface (Mészáros et al., 2007). The core residues at the interface are the hydrophobic residues, generally in the central region of the interface, and play an important role in the interaction. The rim residues are the polar residues, located on the outer edges of the interface. The support residues represent the intersection between the interior and the interface (Levy, 2010).

To examine the efficiency of interactions in different secondary structural elements (SSEs), the average number of external contacts of the interface, core, and rim residues were calculated for each ProS and heterodimer (see Figs. 3.3 A-F).Compared to heterodimers, irregular residues of ProSs have a larger number of contacts than their alpha helices. In the tables 3.1 and 3.2, the P-values of alpha helices and irregulars are shown respectively, for the interface, core, and rim.

3.3.3 Relative ASA (rASA) of secondary structure elements (SSEs)

Figure 2.5 B in chapter 2 showed that the average ΔrASA correlates well with the average number of contacts in ProSs. ΔrASA of each residue is defined by the difference between rASA of the unbound state (rASAm) and that of the bound state (rASAc), and both rASAs are used to define the core, rim and support residues (ΔrASA = rASAm -rASAc) (Levy, 2010).

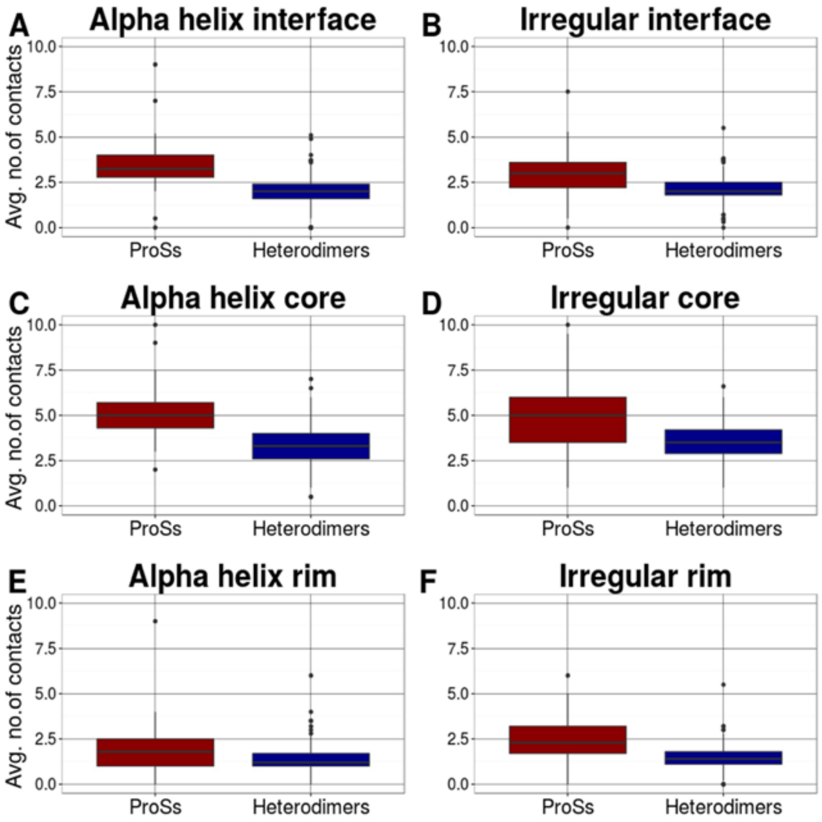

FIGURE 3.3: Interactions of the secondary structure elements (SSEs) in ProSs and heterodimers. Box-plots of the average number of contacts of the alpha helices and irregulars in ProSs and heterodimers at the interface (A and B), core (C and D) and rim (E and F). The distributions of ProSs and heterodimers are colored in red and blue, respectively. The differences between the distributions were evaluated, and the P-values are shown in Table 3.1 and 3.2.

Here, the rASAs of the alpha helices and irregular structures in each ProS and heterodimer at the interface, core and rim were analyzed in detail. In the Figs. 3.4 A-C, D-F, and G-I, the distribution of the average rASAm, rASAc and ΔrASA of ProS alpha helices is shown respectively, for the interface, core, and rim, and compared with those of heterodimers.

Similarly, in the Figs. 3.5 A-C, D-F and G-I, the distribution of the average rASAm, rASAc and ΔrASA of ProS irregulars is shown respectively, for the interface, core, and rim, and compared with those of heterodimers. In both the core and rim, irregular residues of ProSs have a larger rASA in the monomeric state than the heterodimers. The differences are confirmed by a statistical test (See Tables 3.1 and 3.2). The rASA of ProS irregular residues in the monomeric state (rASAm) is large, resulting in a larger ΔrASA that leads to the formation of many contacts.

Contour plots of average rASAm and rASAc of alpha helices and irregular structures are shown in Fig. 3.6 and3.7.

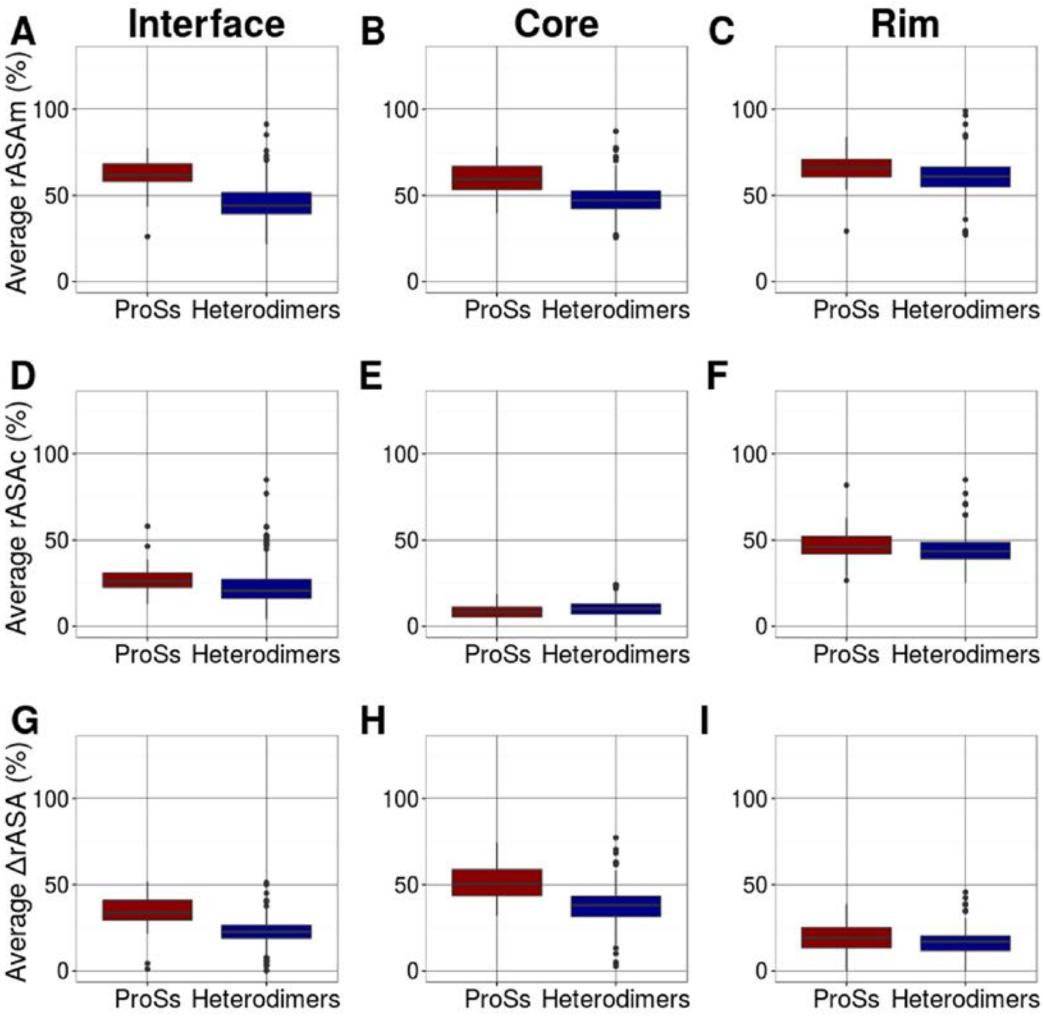

FIGURE 3.4: Average rASAs of alpha helices in ProSs and heterodimers. Average rASAm at the interface (A), core (B) and rim (C).Average rASAc at the interface (D), core (E) and rim (F). Average ΔrASA at the interface (G), core (H) and rim (I). The differences between the distributions were evaluated, and the P-values are shown in Table 3.1.

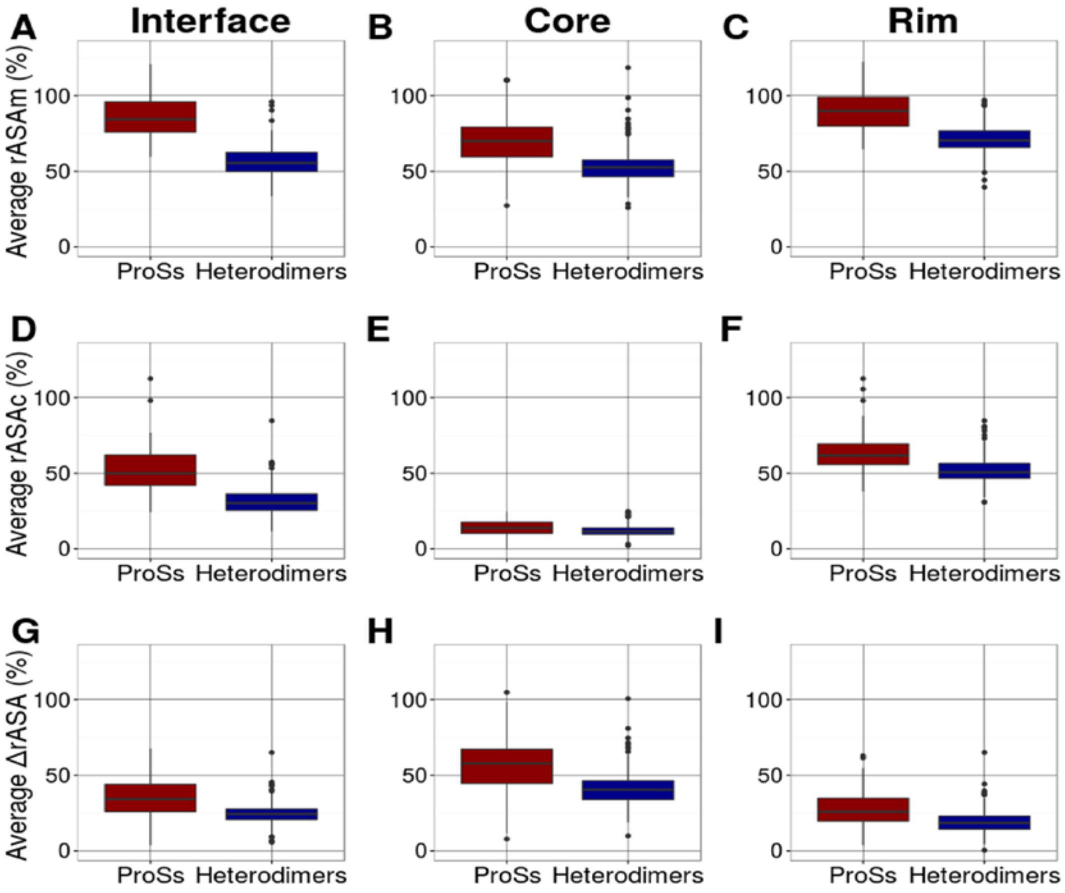

FIGURE 3.5: **Average rASAs of irregular structures in ProSs and heterodimers.** Average rASAm at the interface (A), core (B) and rim (C). Average rASAc at the interface (D), core (E) and rim (F). Average ΔrASA at the interface (G), core (H) and rim (I). The differences between the distributions were evaluated, and the P-values are shown in Table 3.2.

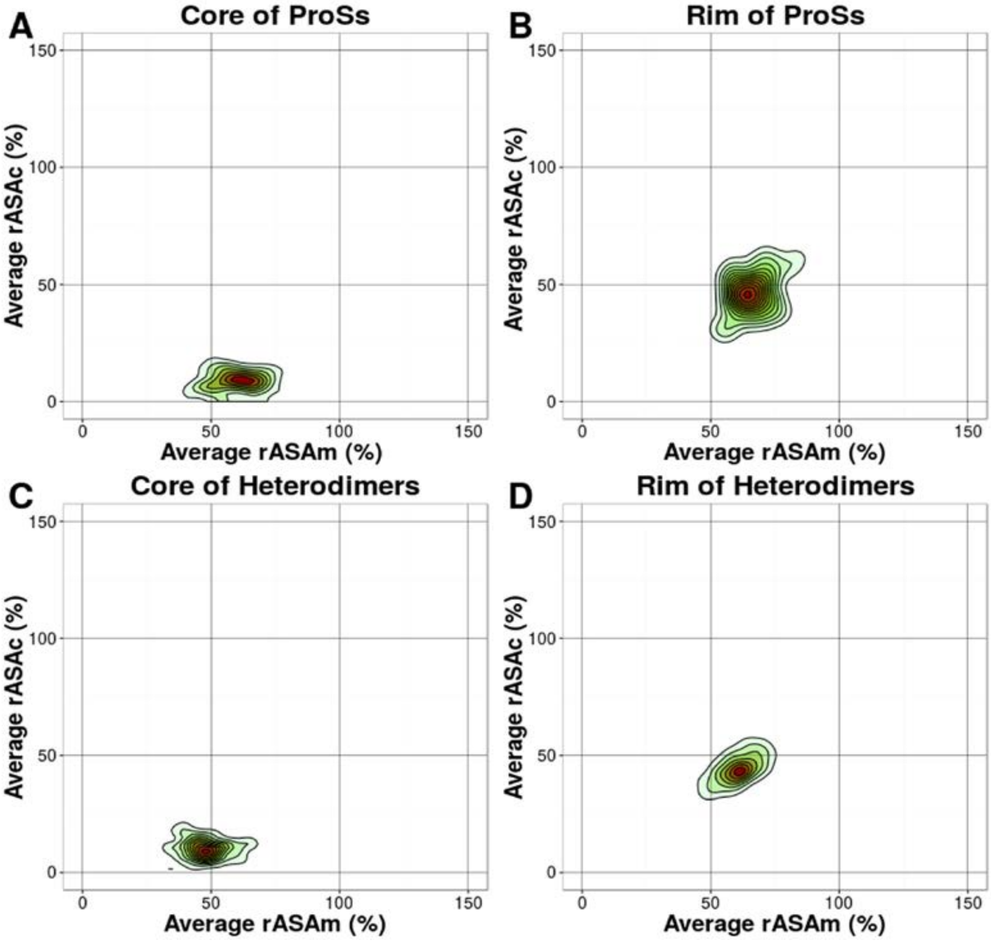

FIGURE 3.6: Contour plots of the average rASAm and average rASAc in alpha helices. (A) Average rASAm vs. average rASAc of the ProS core. (B) Average rASAm vs. average rASAc of the ProS rim. (C) Average rASAm vs. average rASAc of the heterodimer core. (D) Average rASAm vs. average rASAc of the heterodimer rim. The rASAs of each residue in the monomeric and in the complexed states in ProSs and heterodimers were calculated using Naccess (Hubbard and Thornton, 1993). The highest density regions are shown in red, and the lowest density regions are in green.

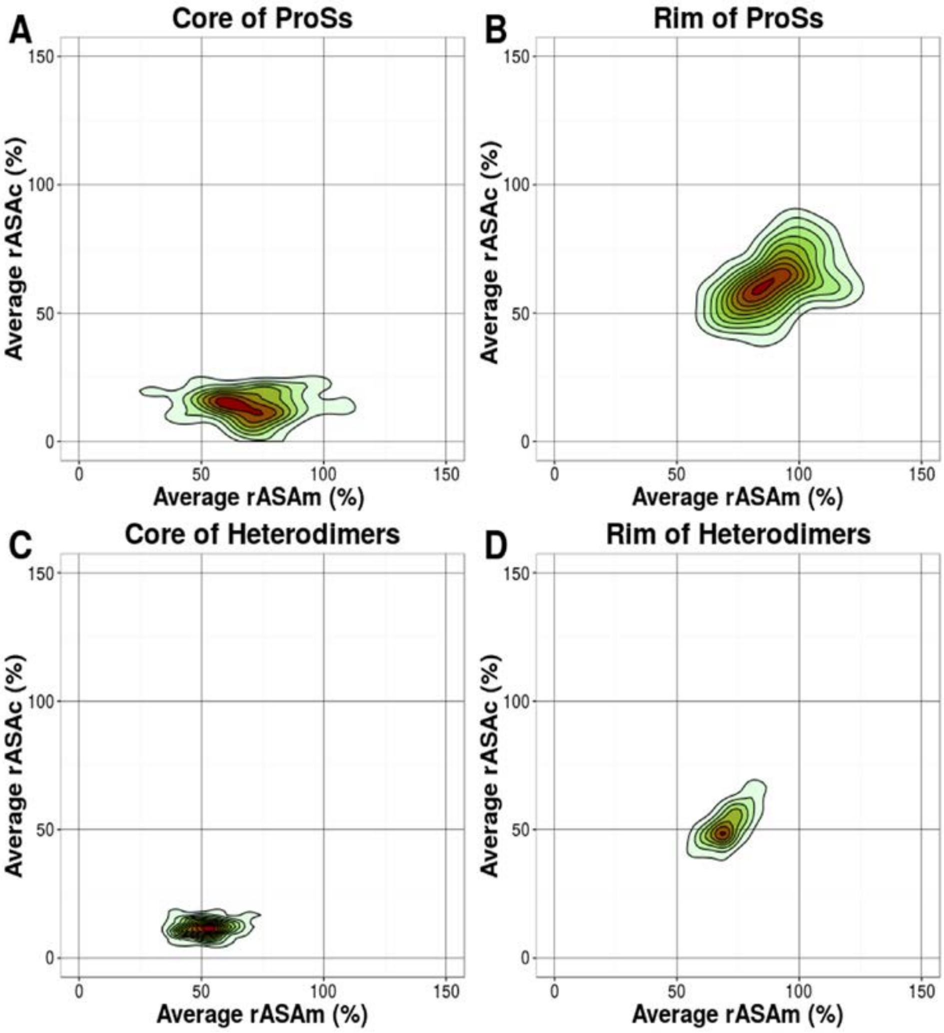

FIGURE 3.7: Contour plots of the average rASAm and average rASAc in irregulars. (A) Average rASAm vs. average rASAc of the ProS core. (B) Average rASAm vs. average rASAc of the ProS rim. (C) Average rASAm vs. average rASAc of the heterodimer core. (D) Average rASAm vs. average rASAc of the heterodimer rim. The rASAs of each residue in the monomeric and in the complexed states in ProSs and heterodimers were calculated using Naccess (Hubbard and Thornton, 1993). The highest density regions are shown in red, and the lowest density regions are ingreen.

Table 3.1: *P*-values of alpha helices in ProSs and heterodimers (using the Wilcoxon rank-sum test)

Features	Places	P-values
Average number of contacts	Interface	1.19e-18
	Core	3.47e-16
	Rim	0.0002
Average rASAm	Interface	3.26e-19
	Core	1.22e-13
	Rim	4.04e-05
Average rASAc	Interface	2.95e-05
	Core	0.008
	Rim	0.044
Average ∆rASA	Interface	1.15e-18
	Core	2.47e-14
	Rim	0.015

Table 3.2: ***P*-values of irregular structures in ProSs and heterodimers (using the Wilcoxon rank-sum test)**

Features	Places	P-values
Average number of contacts	Interface	2.36e-13
	Core	1.89e-09
	Rim	1.19e-16
Average rASAm	Interface	2.15e-44
	Core	4.15e-17
	Rim	1.80e-29
Average rASAc	Interface	3.52e-33
	Core	0.0009
	Rim	4.16e-18
Average ∆rASA	Interface	1.24e-14
	Core	9.79e-13
	Rim	5.21e-12

3.3.4 High and low efficient ProSs

Based on the average number of contacts at the interface, the ProSs were classified into high and low efficient ProSs (See Methods). To examine the properties of high efficient ProSs, several factors, such as average rASAm, average rASAc, average ΔrASA, rate of the interface, rate of the core, rate of the rim, radius of gyration (Rg), and length of the ProSs for each high and low efficient ProS were analyzed. Boxplots of the distributions of high and low efficient ProSs are shown in the Figs. 3.8 A-H. P-values of the high and low efficient ProSs are shown in Table 3.3.

The radius of gyration is used to describe the compactness of a protein, as well as the folding process from the denatured state to the native state (Hong and Lei, 2009; Lobanov, Bogatyreva, and Galzitskaya, 2008). The results show that there is no significant difference between the normalized radiuses of gyration (Rg) of high and low efficient ProSs. Similarly, the factors, such as average rASAc, rate of the interface, rate of the core, rate of the rim, and length of the ProSs are not statistically significant in both high and low efficient ProSs. The reason for this may be the low number of protean segments (ProSs) in the high and low efficient datasets. Interestingly, only the average rASAm and average ΔrASA are statistically significant. This confirms the hypothesis that average rASA in the monomeric state (rASAm) plays a major role in the efficient interactions of ProSs.

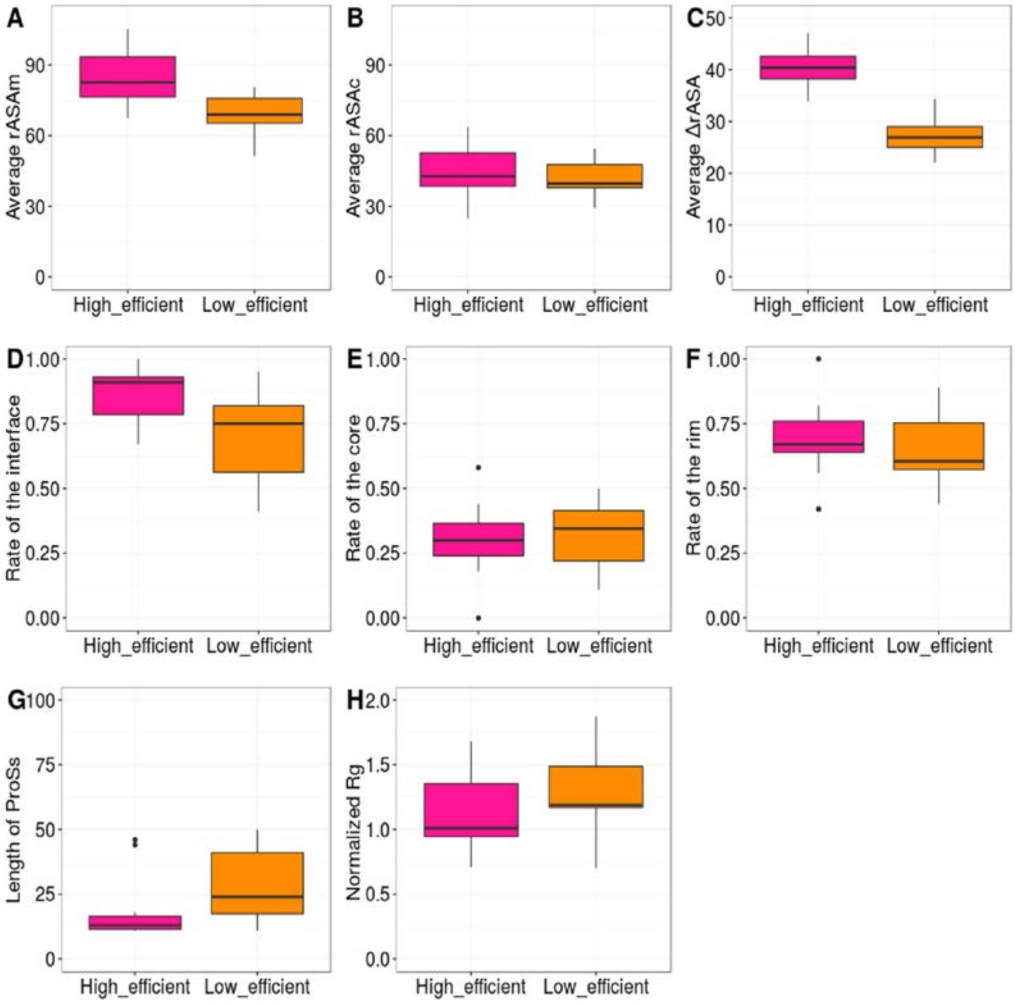

FIGURE 3.8: Box plots of the high and low efficient ProSs. The ProSs are classified into high and low efficient ProSs based on the average number of contacts at the interface. The high efficient ProSs is shown in pink and the low efficient ProS is shown in orange.

Table 3.3: ***P*-values of high and low efficient ProSs (using the Wilcoxon rank- sum test)**

Features	P-values
Average rASAm	0.0007
Average rASAc	0.403
Average ∆rASA	8.97e-07
Rate of the interface	0.028
Rate of the core	0.546
Rate of the rim	0.366
Length of the ProSs	0.02
Normalized Rg	0.228

3.4 Conclusion

The properties of secondary structure elements (SSEs) at the interface, core, and rim of ProSs were analyzed relative to those of heterodimers. The results demonstrate that irregular structures of ProSs and heterodimers are significantly different. Moreover, irregular residues of ProSs have larger number of contacts than their alpha helices. Irregular structures have a larger rASA in the monomeric state (rASAm) that leads to the formation of many contacts inProSs.

Chapter 4

Summary

Summary

This dissertation aims to understand the interaction efficiency of protean segments (ProSs) compared to those of ordered proteins (heterodimers). Protean segments (ProSs) are the short regions in Intrinsically Disordered Proteins (IDPs) that can undergo disorder-to-order transitions upon binding to their protein or nucleic acid partners. Intrinsically disordered proteins (IDPs) are the proteins that lack a fixed three dimensional structure. Intrinsically disordered regions (IDRs) are defined as entire proteins or regions of Intrinsically Disordered proteins (IDPs). Intrinsically Disordered proteins have been shown to be involved in a variety of biological functions.

In this study, we focused on the interface of ProSs and compared to those of heterodimers. Previous studies focused more on the interface of the ProS complexes. To understand the characteristics of ProSs interfaces that are responsible for the interaction efficiency, we classified the interface into core, rim, and support, and analyzed them based on the relative solvent accessible surface area (rASA).

The residues in the interface core are the most buried residues upon protein binding, generally at the central region of the interface, and play an important role in the interaction, like hot spots. The residues in the interface rim are located on the outer edges of the interface that remain partially exposed to the solvent .The support residues play an insignificant role in the interaction, which represents the intersection between the interior and the

interface. Previous studies have shown that the core region is enriched with hydrophobic residues and the rim region is enriched with polar and charged residues.

The relative accessible surface area (rASA) is defined as the total accessible surface area (ASA) of the residues in a protein structure normalized by the ASA of the residues in the most exposed state to a solvent molecule, generally water. The rASAs of each residue, in the monomeric and complex states (rASAm and rASAc, respectively) were computed for ProSs and heterodimers using the program Naccess, which is an implementation of Lee and Richard's algorithm. Residues with rASA less than 25% in the complexed state were assigned to the core and the residues with rASA greater than 25% in the complexed state were assigned to the rim. The remaining residues were assigned to the support.

Compared to the heterodimers, the ProS interface is composed of a small core and a large rim. The distribution of the rates of core and rim is significantly different in ProSs and heterodimers as assessed by the Wilcoxon rank-sum test. Our statement based on the rates slightly different from the results of Wong et al., denoting that the number of residues in the core of the 1D segment (corresponding to ProS) is smaller than that of the 3D complex proteins (heterodimers), but in the rim, the numbers of residues are almost equal.

Meszaros et al. has been indicated that the average number of contacts of ProS at the interface is larger than those of ordered proteins. To examine the efficiency of interactions, we calculated the average number of external contacts of interface residues for each ProS and heterodimer, and compared their distributions. External contacts are defined as the contacts between the proteins and their interaction partners. Our results showed that the ProS interface can be in contact with a larger number of residues of the interaction partners compared with the heterodimer interface, confirming that the ProS interaction is effective.

To analyze the contribution of residues, the average number of contacts was calculated individually for the core and the rim of the interfaces. Our results showed that in both the core and rim cases, the average number of contacts by ProS residues is larger than that by heterodimers. This shows that ProS residues contribute to effective interactions not only through the core, but also through the rim. This indicates the efficiency of the rim region in ProSs. To prove this hypothesis, we ignored the interactions of the core, rim or support individually, and calculated the average number of contacts again. When we took into account the rim contacts and ignored those of the core and support, the average number of contacts of the ProS was different from that of the heterodimer. Only when we ignored the rim contacts, the average number of ProS contacts was almost equal to that of the heterodimer, indicating that the contribution of the rim to interactions is significant, and the interaction mechanism of the rim region should be addressed.

We investigated why the ProS rim is more effective than the rim of heterodimers? For that, we checked the relationship between the average number of contacts and the average ΔrASA which showed a good correlation. We confirmed that when we used ΔrASA instead of ΔASA the same relationship was held. ΔrASA of each residue is defined by the difference between rASA of the monomeric state (rASAm) and that of the complex state (rASAc), and both rASA are used to define the core, rim and support residues. Therefore, analyzing rASA is promising for connecting the feature of the ProS rim with its effective interactions. Our results indicate that when the average number of contacts of ProSs is larger than that of heterodimers, it would be caused by a larger average rASAm of ProS, a smaller average rASAc of ProS, or both.

To examine the structural features, the amino acid compositions of ProSs vs. heterodimers were examined. We computed the Chou–Fasman parameters for core and rim residues. Positive correlations were observed between ProS core vs. heterodimer core, and ProS rim vs. heterodimer rim. Therefore, the amino acid composition cannot explain why the rASAs of the ProS rim are large in the monomeric state.

We computed the Chou–Fasman parameters of the protein interior and IDR using the monomeric proteins in Protein Quaternary Structure Fileserver (PQS) (sequence identity < 25%), and compared them with those of the core and rim .The core and rim residues correlated well with the protein interior and IDR, respectively, indicating that the core residues are hydrophobic and the rim residues are polar.

In summary, the balance between a small core and a large rim, and the large solvent exposure of the rim in the monomeric state, are the key to the disorder-to-order transition and the effective interactions of ProSs. Our study revealed the significance of the rim regions in ProS interactions.

In addition, the characteristics of secondary structure elements (SSEs) at the interface of ProSs relative to those of heterodimers were investigated. The DSSP program (Define Secondary Structure of Proteins) was used to assign the secondary structures in ProSs and heterodimers. The eight types calculated by DSSP were reduced to three, such as alpha helices (H, G and I), beta strands (E) and irregulars (B, S, T and C). The secondary structure distribution of ProSs interface is very different from those of heterodimers. The content of irregular structures and beta strands are the largest difference between ProSs and heterodimers. Alpha helices are almost equally abundant in both data sets. Compared to the heterodimers, ProSs are enriched in irregular residues and depleted in beta strand residues. The results showed that irregular structures of ProSs and heterodimers are significantly different. To examine the efficiency of interactions in different secondary structural elements (SSEs), the average number of external contacts of the interface, core, and rim residues was calculated for each ProS and heterodimer. Compared to heterodimers, irregular residues of ProSs have a larger number of contacts than their alpha helices.

The relative solvent accessible surface areas (rASAs) of the alpha helices and irregular structures in each ProS and heterodimer at the interface, core and rim were analyzed in detail. In both the core and rim, irregular residues of ProSs have a larger rASA in the monomeric state than heterodimers. The rASA of ProS irregular residues in the monomeric state (rASAm) is large, resulting in a larger ΔrASA, which leads to the formation of many contacts.

Furthermore, the ProSs were classified into high and low efficient ProSs based on their average number of contacts at the interface. High and low efficient ProSs were defined as the contacts of ProSs with greater than 4 and less than 2.5, respectively. To examine the properties of high efficient ProSs, several factors, such as average rASAm, average rASAc, average ΔrASA, rate of the interface, rate of the core, rate of the rim, radius of gyration (Rg), and length of the ProSs were analyzed. Interestingly, only the average rASAm and average ΔrASA are statistically significant. The other factors, such as average rASAc, rate of the interface, rate of the core, rate of the rim, and length of the ProSs are insignificant in both high and low efficient ProSs. The reason for this may be the low number of protean segments (ProSs) in the high and low efficient datasets. This confirms the hypothesis that average rASA in the monomeric state (rASAm) plays a major role in the efficient interactions of ProSs.

An unsolved problem in our study is that, we cannot find the reason why

ProSs have larger rASA in the monomeric state? Another disadvantage of this study is the small number of ProSs in the dataset. The properties may differ with a large dataset of ProSs structures.

In this study we revealed the significance of the rim region in ProSs based on rASA. In the future, researchers can try to investigate the other properties of the rim region in IDPs. The revealed characteristics of ProSs (IDP regions) could be used for prediction of binding sites and also help to identify new drug targets. This study will help to find the novel strategies for drug discovery based on IDPs.

Bibliography

Berchanski, Alexander, Boaz Shapira, and Miriam Eisenstein (2004).“Hydrophobic complementarity in protein–protein docking”. In: *PROTEINS: Structure, Function, and Bioinformatics* 56.1, pp.130–142.

Berman, Helen M et al. (2000). “The protein data bank”.In: *Nucleic acids research* 28.1, pp. 235–242.

Cheng, Yugong et al. (2007). “Mining alpha-helix-forming molecular recognition features with cross species sequence alignments”. In: *Biochemistry* 46.47, pp. 13468–13477.

Chothia, Cyrus and Joël Janin (1975). “Principles of protein-protein recognition”. In: *Nature* 256.5520, pp. 705–708.

Chou, Peter Y and Gerald D Fasman (1978). “Empirical predictions of protein conformation”. In: *Annual review of biochemistry* 47.1, pp.251–276.

Davey, Norman E et al. (2012). “Attributes of short linear motifs”. In: *Molecular BioSystems* 8.1, pp. 268–281.

Demchenko, Alexander P (2001). “Recognition between flexible protein molecules: induced and assisted folding”. In: *Journal of molecular recognition* 14.1, pp. 42–61.

Di Domenico, Tomás et al. (2012). “MobiDB: A comprehensive database of intrinsic protein disorder annotations”. In: *Bioinformatics* 28.15, pp. 2080– 2081.

Dinkel, Holger et al. (2013). “The eukaryotic linear motif resource ELM: 10 years and counting”. In: *Nucleic acids research*, gkt1047.

Dunker, A Keith et al. (2001). “Intrinsically disordered protein”. In: *Journal of Molecular Graphics and Modelling* 19.1, pp. 26–59.

Dunker,A Keith et al. (2002). “Intrinsic disorder and protein function”. In:*Biochemistry* 41.21, pp. 6573–6582.

Dyson, H Jane and Peter E Wright (2002). “Coupling of folding and binding for unstructured proteins". In: *Current opinion in structural biology* 12.1, pp. 54–60.

Dyson, H Jane and Peter E Wright (2005). "Intrinsically unstructured proteins and their functions". In: *Nature reviews Molecular cell biology* 6.3, pp. 197–208.

Eisenberg, David and Andrew D McLachlan (1985). "Solvation energy in protein folding and binding." In: *Nature* 319.6050, pp. 199–203.

Fukuchi, Satoshi et al. (2012). "IDEAL: Intrinsically disordered proteins with extensive annotations and literature". In: *Nucleic acids research* 40.D1, pp. D507–D511.

Fukuchi, Satoshi et al. (2014). "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners". In: *Nucleic acids research* 42.D1, pp. D320–D325.

Fuxreiter, Monika et al. (2004). "Preformed structural elements feature in partner recognition by intrinsically unstructured proteins". In: *Journal of molecular biology* 338.5, pp. 1015–1026.

Galea, Charles A et al. (2008). "Regulation of Cell Division by Intrinsically Unstructured Proteins: Intrinsic Flexibility, Modularity, and Signaling Conduits". In: *Biochemistry* 47.29, pp. 7598–7609.

Grant, Barry J et al. (2006). "Bio3d: an R package for the comparative analysis of protein structures". In: *Bioinformatics* 22.21, pp. 2695–2696.

Gsponer, Jörg and M Madan Babu (2009). "The rules of disorder or why disorder rules". In: *Progress in biophysics and molecular biology* 99.2, pp. 94–103.

Guharoy, Mainak and Pinak Chakrabarti (2005). "Conservation and relative importance of residues across protein-protein interfaces". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15447–15452.

Henrick, Kim and Janet M Thornton (1998). "PQS: A protein quaternary structure file server". In: *Trends Biochem Sci.* 23, pp. 358–361.

Heringa, Jaap and Patrick Argos (1991). "Side-chain clusters in protein structures and their role in protein folding". In: *Journal of molecular biology* 220.1, pp. 151–171.

Hobohm, Uwe et al. (1992). "Selection of representative protein data sets". In: *Protein Science* 1.3, pp. 409–417.

Hong, Liu and Jinzhi Lei (2009). "Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity". In: *Journal of Polymer Science Part B: Polymer Physics* 47.2, pp. 207–214.

Hsu, Wei-Lun et al. (2013). "Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding". In: *Protein Science* 22.3, pp. 258–273.

Hubbard, Simon J and Janet M Thornton (1993). "Naccess". In: *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 2.1.

Iakoucheva, Lilia M et al. (2002). "Intrinsic disorder in cell-signaling and cancer-associated proteins". In: *Journal of molecular biology* 323.3, pp. 573– 584.

Iakoucheva, Lilia M et al. (2004). "The importance of intrinsic disorder for protein phosphorylation". In: *Nucleic acids research* 32.3, pp. 1037–1049.

Ihaka, Ross and Robert Gentleman (1996). "R: a language for data analysis and graphics". In: *Journal of computational and graphical statistics* 5.3, pp. 299–314.

Jones, Susan and Janet M Thornton (1996). "Principles of protein-protein in teractions". In: *Proceedings of the National Academy of Sciences* 93.1, pp. 13– 20.

Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22.12, pp. 2577–2637.

Keskin, Ozlem, Buyong Ma, and Ruth Nussinov (2005). "Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues". In: *Journal of molecular biology* 345.5, pp. 1281–1294.

Lee, Byungkook and Frederic M Richards (1971). “The interpretation of protein structures: estimation of static accessibility”. In: *Journal of molecular biology* 55.3, 379–IN4.

Levy, Emmanuel D (2010). “A simple definition of structural regions in proteins and its use in analyzing interface evolution”. In: *Journal of molecular biology* 403.4, pp. 660–670.

Linding, Rune et al. (2003a). “GlobPlot: Exploring protein sequences for globularity and disorder”. In: *Nucleic acids research* 31.13, pp. 3701–3708.

Linding, Rune et al. (2003b). “Protein disorder prediction: implications for structural proteomics”. In: *Structure* 11.11, pp. 1453–1459.

Lobanov, M Yu, NS Bogatyreva, and OV Galzitskaya (2008). “Radius of gyration as an indicator of protein structure compactness”. In: *Molecular Biology* 42.4, pp. 623–628.

Mészáros, Bálint et al. (2007). “Molecular principles of the interactions of disordered proteins”. In: *Journal of molecular biology* 372.2, pp. 549–561.

Minezaki, Yoshiaki et al. (2006). “Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation”. In: *Journal of molecular biology* 359.4, pp.1137–1149.

Mohan, Amrita et al. (2006). “Analysis of molecular recognition features (MoRFs)”. In: *Journal of molecular biology* 362.5, pp. 1043–1059.

Murzin, Alexey G et al. (1995). “SCOP: a structural classification of proteins database for the investigation of sequences and structures”. In: *Journal of molecular biology* 247.4, pp. 536–540.

Nath Jha, Anupam, Saraswathi Vishveshwara, and Jayanth R Banavar (2010). “Amino acid interaction preferences in proteins”. In: *Protein Science* 19.3, pp. 603–616.

Oates, Matt E et al. (2013). “D2P2: database of disordered protein predictions”. In: *Nucleic acids research* 41.D1, pp. D508–D516

Oldfield, Christopher J et al. (2005). “Coupled folding and binding with *a*-

helix-forming molecular recognition elements”. In: *Biochemistry* 44.37, pp. 12454– 12470.

Ooi, Tatsuo et al. (1987). “Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides”. In: *Proceedings of the National Academy of Sciences* 84.10, pp. 3086–3090.

Prilusky, Jaime et al. (2005). “Foldindex: A simple tool to predict whether a given protein sequence is intrinsically unfolded”. In: *Bioinformatics* 21.16, pp. 3435–3438.

Racine, Jeffrey S (2012). “RStudio: A Platform-Independent IDE for R and Sweave”. In: *Journal of Applied Econometrics* 27.1, pp. 167–172.

Reichmann, Dana et al. (2007). “The molecular architecture of protein–protein binding sites”. In: *Current opinion in structural biology* 17.1, pp.67–76.

Romero, Pedro et al. (2001). “Sequence complexity of disordered protein”. In:*Proteins: Structure, Function, and Bioinformatics* 42.1, pp. 38–48.

Rose, George D et al. (1985). “Hydrophobicity of amino acid residues in globular proteins”. In: *Science* 229.4716, pp.834–838.

Sickmeier, Megan et al. (2007). “DisProt: the database of disordered proteins”. In: *Nucleic acids research* 35.suppl 1, pp. D786–D793.

Thompson, Julie D, Desmond G Higgins, and Toby J Gibson (1994). “CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. In: *Nucleic acids research* 22.22, pp.4673–4680.

Tompa, Peter et al. (2009). “Close encounters of the third kind: disordered domains and the interactions of proteins”. In: *Bioessays* 31.3, pp. 328–335.

Uversky, Vladimir N (2012).“Intrinsically disordered proteins and novel strategies for drug discovery”. In: *Expert opinion on drug discovery* 7.6, pp.475–488.

Uversky, Vladimir N (2014). “Introduction to intrinsically disordered proteins (IDPs)”. In: *Chem- ical reviews* 114.13, pp.6557–6560.

Uversky, Vladimir N, Christopher J Oldfield, and A Keith Dunker (2008). “Intrinsically disordered proteins in human diseases: introducing the D2 concept”. In: *Annu. Rev. Biophys.* 37, pp. 215–246.

Uversky, Vladimir N et al. (2014). “Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases”. In: *Chem. Rev* 114.13, pp. 6844–6879.

Vacic, Vladimir et al. (2007). “Characterization of molecular recognition features, MoRFs, and their binding partners”. In: *Journal of proteome research* 6.6, pp. 2351–2366.

Van Der Lee, Robin et al. (2014). “Classification of intrinsically disordered regions and proteins”. In: *Chemical reviews* 114.13, pp. 6589–6631.

Varadi, Mihaly et al. (2014). ”pE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins”. In: *Nucleic acids research* 42.D1, pp. D326–D335.

Vucetic, Slobodan et al. (2003). “Flavors of protein disorder”. In: *Proteins: Structure, Function, and Bioinformatics* 52.4, pp. 573–584.

Ward, Jonathan J et al. (2004a).“Prediction and functional analysis of native disorder in proteins from the three kingdoms of life”. In: *Journal of molecular biology* 337.3, pp. 635–645.

Ward, Jonathan J et al. (2004b). “The DISOPRED server for the prediction of protein disorder”. In: *Bioinformatics* 20.13, pp. 2138–2139.

Wong, Eric TC, Dokyun Na, and Jörg Gsponer (2013). “On the importance of polar interactions for complexes containing intrinsically disordered proteins”. In: *PLoS Comput Biol* 9.8, e1003192.

Wright, Peter E and H Jane Dyson (1999). “Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm”. In: *Journal of molecular biology* 293.2, pp. 321–331.

Wright, Peter E and H Jane Dyson (2015). “Intrinsically disordered proteins in cellular signalling and regulation”. In: *Nature Reviews Molecular Cell Biology*

16.1, pp. 18–29.

Xue, Bin et al. (2010). "PONDR-FIT: A meta-predictor of intrinsically disordered amino acids". In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1804.4, pp. 996–1010.

Yan, Changhui et al. (2008). "Characterization of protein–protein interfaces". In: *The protein journal* 27.1, pp. 59–70.

Young,L, RL Jernigan, and DG Covell(1994)."A role for surface hydrophobicity in protein-protein recognition". In: *Protein Science* 3.5, pp. 717–729.

Zhou, Pei et al. (2001). "Solution structure of DFF40 and DFF45 N-terminal domain complex and mutual chaperone activity of DFF40 and DFF45". In: *Proceedings of the National Academy of Sciences* 98.11, pp. 6051–6055.