# Protein-protein interactions: a structural bioinformatics approach

**Sowmya Gopichandran**

*M.Sc. Biotechnology (By Research),*

*AIMST University, Malaysia*

A thesis submitted in fulfilment of the requirements for the degree of

**Doctor of Philosophy**

June 2015

Department of Chemistry and Biomolecular Sciences

Macquarie University, Sydney, Australia

***DEDICATED TO SCIENCE,***

*enlightening us*

# DECLARATION

I certify that this thesis entitled "Protein-protein interactions: a structural bioinformatics approach" is a bonafide record of research work carried out by me under the guidance of Professor Shoba Ranganathan during the years 2011-2015 for the degree of Doctor of Philosophy. The results presented in this thesis have not previously formed the basis for award of any degree, fellowship or other recognition. The particulars given in the thesis are true to the best of my knowledge.

Sowmya Gopichandran

June 2015

# Acknowledgements

It is my pleasure to thank the following people who made this thesis possible through their continuous encouragement, motivation and support:

## Professional

- My supervisor, Prof Shoba Ranganathan, for her support and guidance during this work. I am very thankful for her patience, motivation, enthusiasm and intellectual support and also for giving an opportunity to be a part of her group.
- My co-supervisor, Prof Mark Baker for his support through the first two years of my PhD tenure.
- Prof Helena Nevalainen for her support and counsel during my PhD.
- Prof Pandjassarame Kangueane, my ex-supervisor for his continued support, encouragement and valuable suggestions throughout my PhD.
- Ms. Catherine Wong, Ms. Michelle Kang, Ms. Jane Yang, for administrative support.
- Macquarie University, for the award of iMQRES research scholarship for pursuing a PhD and PGRF funding for attending the ISCB 2014 conference in Boston, MA, USA.
- Past and current colleagues: Dr Gagan Garg, Mr Mohammad Islam, Dr Abidali Mohamedali, Mr Ari Haridianto, Ms Ishmam Nawar and Ms Criselda Fernandes.
- I take this opportunity to express gratitude to all of the Department faculty members for their help and support.
- I also place on record, my sense of gratitude to one and all who directly or indirectly, have lent their hand in this venture.

## Personal

- I am extremely grateful to my parents (Mrs KR Suseelamba and Mr E Gopichandran), brother (Mr G Madhusudhan) and grandmother (Mrs KR Jayalaxmamma) for their endless love, unceasing support, and encouragement.
- Hearty thanks to my fiancé, Tarun for his constant encouragement and moral support.
- Special thanks to my friend Jashanpreet Kaur for her love and support during tough times.
- I would also like to thank my relatives in Sydney for their help and support.
- Thanks to all my dear friends in Australia and abroad, and my relatives for their love and encouragement.
- Last, but not the least, I am eternally grateful to God for all the wonderful things he has given me, including good health and wellbeing that were necessary to complete this PhD.

# Table of Contents

## List of Abbreviations

| | |
|---|---|
| **2S** | 2-state without any intermediate |
| **3D** | Three dimension |
| **3did** | 3D Interacting Domain |
| **3SDI** | 3-state with dimer intermediate |
| **3SMI** | 3-state with monomer intermediate |
| **AA** | Amino Acid |
| **AD** | Activation domain |
| **ASA** | Accessible surface area |
| **B/2** | Interface Area |
| **BE** | Binding energy |
| **BID** | Binding Interface Database |
| **BIND** | Biomolecular Interaction Network Database |
| **BioGRID** | Biological General Repository for Interaction Datasets |
| **BLAST** | Basic Local Alignment Search Tool |
| **CAPRI** | Critical Assessment of Predicted Interactions |
| **CATH** | Protein Structure Classification Database |
| **CBP** | Calmodulin-binding peptide |
| **cDNA** | complementary Deoxyribonucleic acid |
| **Co-IP** | Co-immunoprecipitation |
| **cons-PPISP** | consensus neural-network Protein-Protein Interaction Site Predictor |
| **CRC** | Colorectal cancer |
| **CYGD** | Comprehensive yeast genome database |
| **DBD** | DNA-binding domain |
| **DFHR** | Dihydrofolate reductase |
| **DI** | Domain I |
| **DII** | Domain II |
| **DIII** | Domain III |

| | |
|---|---|
| **DIP** | Database of Interacting Proteins |
| **DNA** | Deoxyribonucleic acid |
| **DOMINE** | Database of domain–peptide interactions |
| **EGTA** | Ethylene glycol tetra acetic acid |
| **ERK** | Extracellular signal-regulated kinases |
| **GAL4** | GALactose metabolism |
| **GO** | Gene Ontology |
| **GST** | Glutathione S-transferase |
| **H-bond** | Hydrogen bond |
| **H-bonds** | Hydrogen bonds |
| **HPID** | Human Protein Interaction Database |
| **HPRD** | Human Protein Reference Database |
| **ICBS** | Inter-Chain Beta-Sheets |
| **IgG** | Immunoglobulin G |
| **IMEx** | International Molecular Exchange Consortium |
| **InterPreTS** | Interaction Prediction through Tertiary Structure |
| **KFC** | Knowledge-based FADE and Contacts |
| **L1** | Linker 1 |
| **L2** | Linker 2 |
| **LAP** | Latency associated peptide |
| **LU** | Ly6/uPAR/α-neurotoxin-like domains |
| **MAP** | Mitogen-activated protein kinase |
| **MAPPIS** | Multiple Alignment of Protein-Protein InterfaceS |
| **meta-PPISP** | meta web server for protein-protein interaction site prediction |
| **MIMIx** | Molecular Interaction Experiment |
| **MINT** | Molecular INTeraction Database |
| **MIPS** | Munich Information Centre for Protein Sequences |
| **MMP** | Matrix metalloproteinase production |

| | |
|---|---|
| **mRNA** | messenger RNA |
| **NGS** | Next generation sequencing |
| **NIH** | National Institutes of Health |
| **NLM** | National Library of Medicine |
| **OMIM** | Online Mendalian Inheritance in Man |
| **OPHID** | Online predicted human interaction database |
| **ORF** | Open reading frame |
| **P%-NP%** | Interface property abundance |
| **PCA** | Protein-fragment complementation assays |
| **PCRPi-W** | Presaging critical residues in protein interfaces-web server |
| **PDB** | Protein Data Bank |
| **Pfam** | Database of protein families |
| **PIC** | Protein Interactions Calculator |
| **PIPS** | Human protein-protein interaction database |
| **PMC** | PubMed Central |
| **PMID** | PubMed identifier |
| **PPI** | Protein-protein interaction |
| **PQS** | Protein Quaternary Structure file server |
| **PredHS** | Prediction of Hot Spots by Using Structural Neighbourhood Properties |
| **PRICE** | PRotein Interface Conservation and Energetics |
| **PSI-MI** | Proteomics Standards Initiative-Molecular Interactions |
| **PTM** | Post-translation modification |
| **R-G-D** | Arginine-glycine-aspartate |
| **RMSD** | Root mean square deviation |
| **r-value** | Pearson correlation coefficient value |
| **SCOP** | Structural Classification of Proteins |
| **SCOPPI** | Structural Classification of Protein-Protein Interfaces |
| **SCOWLP** | Structural Characterisation of Water, Ligands and Protein |

| | |
|---|---|
| **SDS-PAGE** | Sodium dodecyl sulphate-polyacrylamide gel electrophoresis |
| **SNAPPI** | Structures, iNterfaces and Alignments for Protein-Protein Interactions |
| **STRING** | Search Tool for the Retrieval of Interacting Genes/Proteins |
| **TAP** | Tandem affinity purification |
| **TEV** | Tobacco etch virus |
| **TGF-β1** | Transforming Growth Factor-β1 |
| **uPA** | Urokinase plasminogen activator |
| **uPAR** | Urokinase plasminogen activator receptor |
| **WHISCY** | What information does surface conservation yield |
| **Y2H** | Yeast two-hybrid system |
| **ΔASA** | Change in accessible surface area |
| $\Delta^{i}G$ | Solvation free energy gain upon interface formation |
| $\Delta\Delta G_{el}$ | Interface electrostatic energy |

## List of Tables

## List of Figures

## List of Publications included in this thesis

The following papers are presented in this thesis and are referred to from this point onwards as listed in respective sections of the thesis, with my contributions to each paper:

***Publications***

[1] **Sowmya G,** Ranganathan S. Protein-protein Interactions and Prediction: A Comprehensive Overview (2013), **Protein Peptide Letters** 21:779-789.
Contribution to (i) concept: GS 80%, SR 20%; (ii) data gathering: GS 100%; and (iii) writing: GS 80%, SR 20%

[2] **Sowmya G**, Ranganathan S. Discrete structural features among interface residue-level classes (2015), **BMC Bioinformatics** ***(in press)***.
Contribution to (i) concept: GS 80%, SR 20%; (ii) data gathering: GS 100%; (iii) data analysis: GS 80%, SR 20% and (iii) writing: GS 70%, SR 30%

[3] **Sowmya G**, Breen EJ, Ranganathan S. Linking structural properties of protein complexes and biological function (2015), **Protein Science** 24:1486-1494.
Contribution to (i) concept: GS 80%, SR 20%; (ii) data gathering: GS 100%; (iii) data analysis: GS 50%, EJB 40%, SR 10% and (iii) writing: GS 70%, SR 30%

[4] **Sowmya G,** Khan JM, Anand S, Ahn SB, Baker MS, Ranganathan S. A site for direct integrin αvβ6•uPAR interaction from structural modeling and docking (2014), **Journal of Structural Biology** 185:327-335.
Contribution to (i) concept: GS 30%, JMK 20%, SA 5%, SBA 5%, MSB 10%, SR 30%; (ii) data gathering: GS 70%, JMK 30%; (iii) data analysis: GS 40%, JMK 30%, SR 30% and (iii) writing: GS 70%, JMK 15% SR 15%

***Appendix - I***

[5] Ahn SB, Mohamedali A, Anand S, Cheruku HR, Birch D, **Sowmya G**, Cantor D, Ranganathan S, Inglis DW, Frank R, Agrez M, Nice EC, Baker MS. Characterisation of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics (2014), **Journal of Proteome Research** 13:5956-5964.

### *Oral presentations based on thesis work*

[1] **Sowmya G** and Ranganathan S. A site for direct integrin αvβ6•uPAR interaction from structural modeling and docking. Australian School of Advanced Medicine (ASAM), Macquarie University, Sydney, Australia, Feb 2014 *(publication 4)*

### *Poster presentations*

[1] **Sowmya G** and Ranganathan S. A site for direct integrin αvβ6•uPAR interaction from structural modeling and docking. InCoB 2014 Conference, Sydney, Australia, Aug 2014 *(publication 4)*

[2] **Sowmya G** and Ranganathan S. A site for direct integrin αvβ6•uPAR interaction from structural modeling and docking. ISCB 2014 and 3D-Sig 2014 Conferences, Boston, USA, Jul 2014 *(publication 4)*

[3] **Sowmya G** and Ranganathan S. A site for direct integrin αvβ6•uPAR interaction from structural modeling and docking. Sydney Bioinformatics Research Symposium, Garvan Institute, Sydney, Australia, Nov 2013 *(publication 4)*

### *Awards received as PhD student*

[1] **Macquarie University Postgraduate Research Fund (PGRF)**, 2014 (awarded A$ 5000 for International conference travel).

[2] **Deputy Vice-Chancellor (Research) Commendations,** 2014 (awarded additional A$ 500 in recognition of outstanding PGRF applications which demonstrate a deep understanding of and commitment to the candidate's chosen field of study, and communication of these with skill and professionalism).

[3] **Outstanding fast-forward poster prize**, Sydney Bioinformatics Research Symposium, 2013 (Presenters had exactly 120 seconds to pitch their work to the audience).

[4] **International Macquarie University Research Scholarship** (iMQRES) holder, 2011-2015.

# Abstract

Molecular function in cellular processes is governed by protein-protein interactions (PPIs). With the exponential growth of PPI in the drug discovery field, understanding the key principles governing PPI is of immense current interest. Investigation of protein interfaces of known complexes is an important step towards understanding the molecular basis of PPIs. The overall objective of this thesis is to study known PPI complexes from the Protein Data Bank (PDB) using computational tools to capture their driving force and relate structural interface features to their biological functions. PPI features, analysis data and conclusions drawn are documented to facilitate prediction of interaction sites and partners and also to facilitate prediction of potential protein function of novel complexes.

The interface features were analysed for all non-redundant protein heterodimers (278) in the PDB. The relative interface-surface polarities of each complex in the dataset were estimated to understand predominant forces driving binding. Structural analysis revealed two classes of interfaces - class A with less polar residues and class B with more polar residues, at the interface than the rest of the surface. Five distinguishing features (interface area, interface property abundance, interface charged residues, solvation free energy gain, binding energy) among these classes were identified. These results verify the need for classification of complexes based on residue-level properties in determining the features driving binding. Also, all functional categories are represented in the interface classes. This led to the study on relating structural features to their biological functions.

PPIs are essential for catalysis, regulation, assembly, immunity and inhibition in a cell. However, it is unclear whether structural features can define protein functionality. Therefore, analysis of non-redundant protein complexes has been carried out to determine the structural basis for functional preferences. Structural interface of each complex has been characterized using a range of physico-chemical properties. The dataset is grouped using known function for molecular preferences. Five interface features (interface area, interface property abundance, hydrogen bonds, salt bridges, solvation free energy gain, and binding energy) are observed to be significantly different among functional groups.

Preliminary application of using PPIs for the characterisation of protein interfaces in integrin αvβ6 heterodimer and its interactions with other proteins especially urokinase plasminogen activator receptor (uPAR) is carried out. The integrin αvβ6•uPAR interaction promotes cancer progression. Therefore, a comprehensive analysis of αvβ6 using modelling data and docking simulations helped gain insights into binding of αvβ6 with uPAR suggesting an interaction site. These results provide preliminary evidence for potential targets in cancer therapies.

In conclusion, the work presented in this thesis investigates interface features of known protein complexes to gain insights into the binding principles of PPIs. Structural analysis of heterodimer dataset and grouping complexes based on interface classes and functional groups lead to the identification of discriminatory features amongst these groups. Incorporation of these combinatorial features is necessary to develop models for PPI prediction and analysis, and also in utilizing PPI information for the prediction of potential functions in future studies. Novel observations using modelling and docking data to obtain significant information on key PPIs (involved in cancer) are discussed.

# Chapter 1: Introduction

## 1.1 Overview

Proteins are essential working molecules of a biological system participating in virtually every process within the living cell. They are polypeptide macromolecules with finite sequences. A polypeptide is a linear chain of amino acid residues bonded together by peptide bonds. Short polypeptides comprising of 20-30 amino acid residues are called peptides or oligopeptides. Proteins differ primarily in their amino acid sequences. The amino acid sequence in a protein is prescribed by the nucleotide sequence of their genes and results in folding of the protein into unique three-dimensional structures for specific activity [10-12]. Protein biosynthesis occurs during the translation of mRNA into polypeptide chains on ribosomes [13].

During or after protein synthesis, the residues are often chemically modified by posttranslational modification altering physicochemical properties, folding, proteolytic cleavage, stability and function of proteins [14]. Proteins thus formed exist for a certain period of time and are then broken down and replaced through a process called protein turnover [15, 16]. The turnover rate for individual proteins is different. The lifespan of a protein is measured in terms of its half-life and varies widely with an average lifespan of 1-2 days in mammalian cells [17, 18]. Protein degradation is inversely proportional to the stability of proteins with misfolded or unfolded proteins degrading more rapidly. Many proteins bind to another protein or multiple proteins, organic and inorganic compounds, metals, sugars, fatty acids and nucleic acids in a cellular system [15].

Protein-protein interactions (PPIs) form the central basis for complex biological networks in a cell. These interactions play a key role in the fields of systems biology, functional genomics and drug design [19-21]. Specific interactions between two or more protein molecules lead to various functions such as catalysis, regulation, signalling, immunity and inhibition [22-24]. Advancements in experimental procedures have led to the determination of PPIs within the genome [25-28]. These experiments could however be laborious and time-consuming with inclusion of false positives that are not necessarily associated *in vivo* [29], thereby posing a need for progress in computational methods in determining PPIs [30].

A stable interface is often formed between the interacting proteins to achieve a particular function. Protein-protein interfaces are extensively analysed for PPI features as a crucial step towards deciphering the binding principles and functionality of proteins [23, 31-33]. These interface features have been investigated over several decades using sequence and structure information [6, 22-24, 30-49]. The interacting residues are characterized based on physical and chemical features, based on their strengths of interactions in different groups of complexes [50]. Several advances in the analyses of the interface residues give insights into the significance of PPI prediction [47, 51-60]. Moreover, the chemical properties of interacting residues gives useful information for various applications such as design of drugs that target such interactions, design of mutants for experimental verifications of interactions, construction of cellular network maps and also in the prediction of binding partners through sequence information, docking procedures and homology modeling [30]. Therefore, a study on the interactions of proteins in known complexes is the key to understanding cell machinery.

In this thesis, a comprehensive literature review on PPIs and the key resources available for the study of PPIs has been carried out (Publication 1). A comprehensive structural analysis of an updated dataset of heterodimer complexes to identify distinguishing PPI features among various classes and thereby understand the structure-function relationship between interacting proteins is carried out. Based on these results, an application of PPI study for the abrogation of colon cancer progression is discussed. The specific aims of this thesis and how they have been addressed forms the rest of the thesis, followed by conclusions and future direction.

At the outset, the various experimental procedures for the determination of PPIs are described. This is followed by the computational methods in determining PPIs, the key databases archiving this information and the current trends involved in studying PPIs is discussed. The key datasets of PPIs created/collected by various groups and the deterministic PPI features known to govern PPIs are then presented. The various protein interface databases and protein interaction characterisation and prediction tools/servers are documented. This is followed by an introduction to integrins. From this, the objectives for the thesis have been set out. A comprehensive overview on PPIs is presented in Publication 1.

## 1.2 EXPERIMENTAL METHODS FOR THE DETERMINATION OF PPIs

Recent progress in the field of proteomics has led to the development of a number of powerful methods for the determination of PPIs. Many new methods are now available for identification and characterisation of PPIs. An advance in techniques such as mass spectrometry has helped identify individual proteins and also characterise biological assemblies [61]. Experiment techniques are based on the qualitative aspects of PPI with less/no information on quantitative determination of PPIs [62]. New methods in proteomics are being developed for the quantitative aspects of PPI.

Although various advances have been made to experimental studies for the determination of PPIs, these state-of-the-art methods are expensive and not available to many labs around the world. Moreover, experimental techniques are laborious and time-consuming with the addition of a high false-positive rate [63]. Experimental methods are also known to generate large amount of data and these need to be statistically verified for possible interpretations. Furthermore, the number of PPIs determined by experimental methods in an entire species is often underestimated [64-66]. Therefore, computational approaches complement experimental methods in narrowing and prioritising the data for effective determinations of PPIs.

Common experimental methods used to identify and verify PPIs are discussed below with advantages and disadvantages of each method.

### 1.2.1 Protein-fragment complementation assays

Protein-fragment complementation assays, or PCAs, are a family of assays that provide a direct method for the identification of PPIs in a living cell. PCAs are widely used in the field of proteomics. In this method, the two proteins of interest, referred to as “bait” and “prey” are covalently linked to fragments of a third reporter protein. Upon interaction of the two proteins (“bait” and “prey”), fragments of the reporter protein come in close proximity and form a functional reporter. The activity of the functional receptor protein is then measured (Figure 1.1). Proteins that can be split and reconstituted by the interacting proteins, can be used as receptor proteins, such as dihydrofolate reductase (DHFR) [67], yeast GAL4 [68], β-lactamase [69], luciferase [70] and Ubiquitin [71] proteins. This principle is the basis for yeast two-hybrid system.

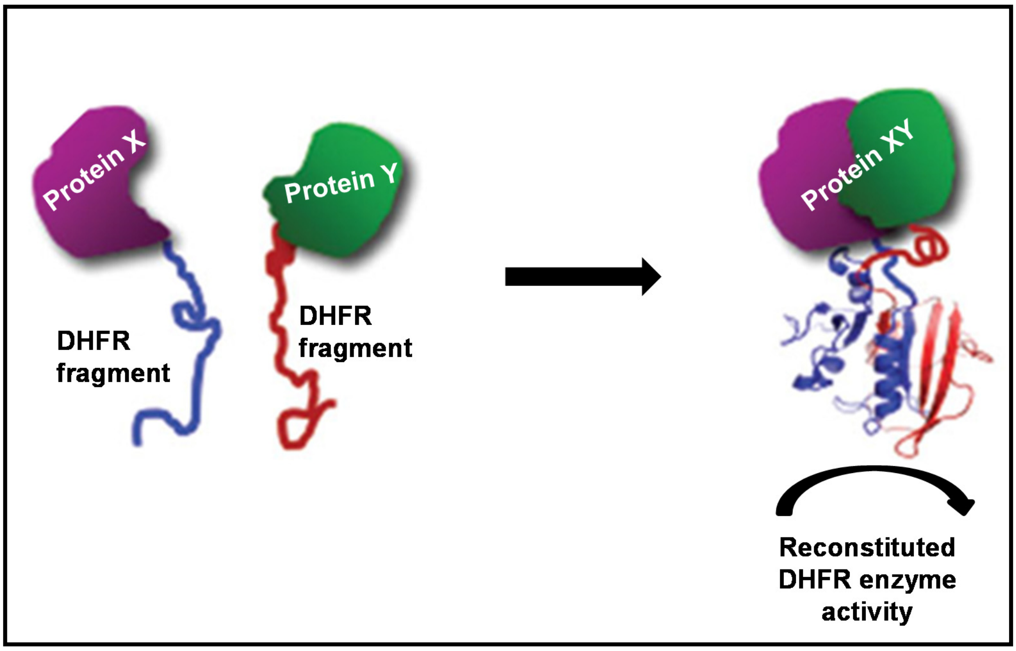


**Figure 1.1: Illustration of protein-fragment complementation assay**. The two proteins of interest (proteins X and Y) are fused to complementary fragment of receptor protein, DHFR. If the two proteins interact, the receptor fragments are brought together folding into native structure thereby reconstituting its activity. Adapted from Remy *et al.,* 2007 [1].

The classic yeast two-hybrid system (Y2H) was developed by Field and Song in 1989 [68], using the properties of GAL4 protein, a transcriptional factor of the yeast *S. cerevisiae.* The GAL4 protein, acts as the reporter protein while the DNA-binding domain (DBD) and the activation domain (AD) act as "bait" and "prey". When the two domains are linked into their functional form, the GAL4 protein activates the expression of the reporter gene and ceases capability of activation while they are split. The proteins of interest are fused to DBD and AD. The interaction of proteins fused to DBD and AD reconstitutes functional form of the domains and thereby allows GAL4 protein to activate the expression of reporter gene. Hence, the activity of GAL4 protein enables effective determination of true interactions between two proteins.

The Y2H approach has disadvantages such as inclusion of high false positive rates, limitation to interactions within the nucleus barely accessing proteins that are anchored to or integrated into the plasma membrane [25]. Therefore, the two-hybrid approach is extended by split receptor techniques, for example, split ubiquitin protein technique [72] is used to include membrane protein interactions. The PCAs have been extended over the years for wide-range screening of protein interactions in different components of the cell and in different organisms and also to reduce false positives [73-75].

### 1.2.2 Affinity purification methods

The principle behind affinity purification method is the use of an affinity-tag fused to specific proteins which interact *in vivo* and are preserved during biochemical purification steps *in vitro*. Widely used affinity purification techniques include Tandem Affinity Purification (TAP) method, GST pull-down technique and Co-immunoprecipitation (Co-IP) augmented with mass spectrometry for protein identification.

#### *1.2.2.1 Tandem Affinity Purification (TAP) method*

The TAP method is one of the best known methods to purify protein complexes and study protein-protein interactions. The TAP tag comprises of a calmodulin-binding peptide (CBP), tobacco etch virus (TEV) protease and protein A, which binds to immobilized Immunoglobulin G (IgG).

In this method, the protein of interest is fused to a TAP tag at its C-terminal and allowed to express and fuse to its targets in a cell line. The TAP tag is cleaved by the enzyme, TEV, which minimizes the cleavage of bait proteins and/or its associated proteins [61]. The protein of interest in then washed through two affinity columns and studied for binding partners (Figure 1.2).

Although TAP tag method is sensitive and selective with quantitative determination of PPI *in vivo*, it has some disadvantages. Tag method could lose transient PPIs during purification steps [76]. They may also lose low-abundance of binding proteins. Moreover, the tag fused to a protein may obscure binding of new proteins. In addition, tags may affect protein expression levels and there is a possibility for the TEV protease enzyme to cleave the proteins, though highly unlikely given the specificity of TEV protease binding. Then again, the tag may also not be adequately exposed to the affinity beads, thereby skewing the results. Several addition techniques such as the GS-TAP-tag methods [77] are more advantageous than other TAP methods such as yeast TAP tag method [78].

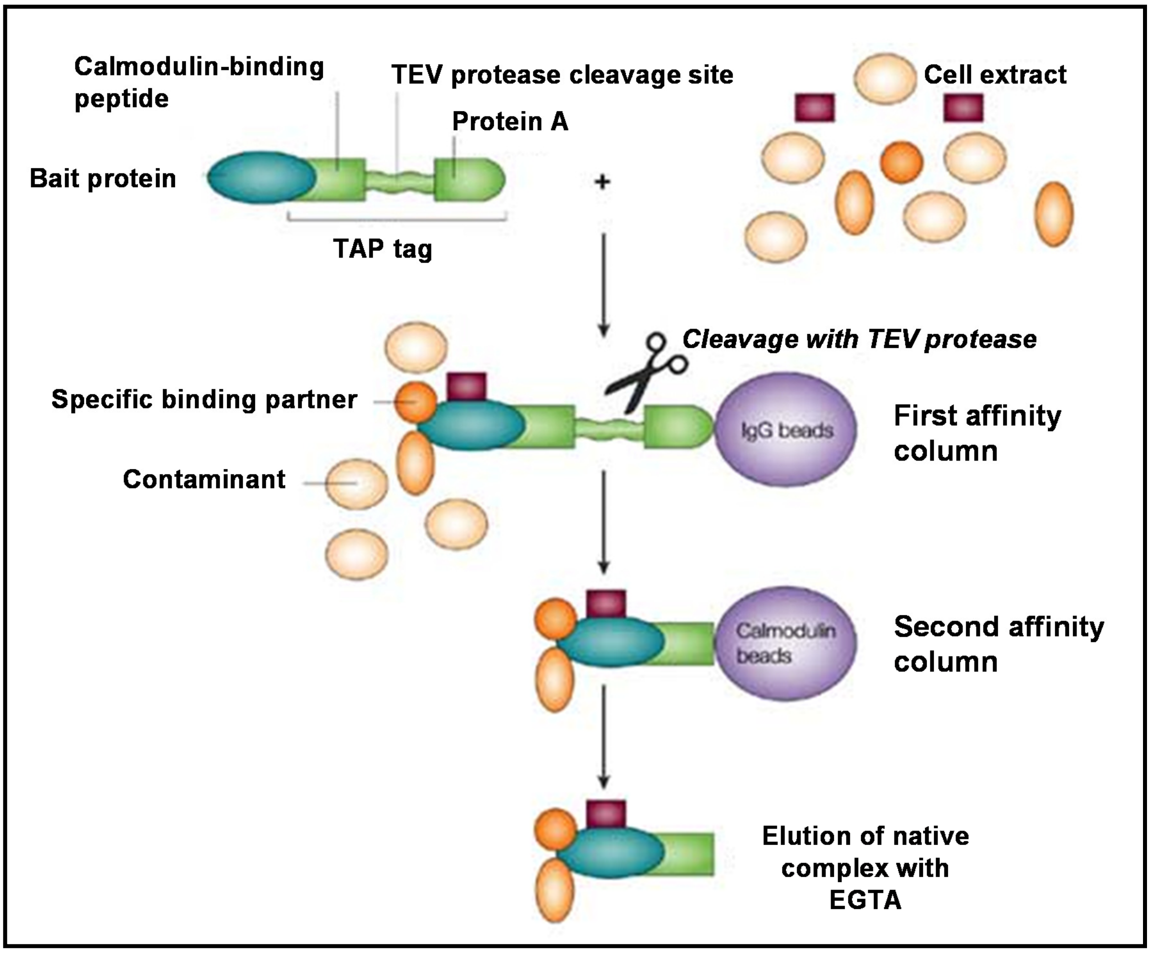


**Figure 1.2: Illustration of the TAP tag method.** The TAP consists of a calmodulin-binding peptide (CBP), tobacco etch virus (TEV) protease and protein A, which binds to immobilized Immunoglobulin G (IgG). Cells containing TAP-tagged proteins are generated and are then extracted under mild conditions and TAP is performed. The first affinity column consists of IgG beads. TEV protease enzyme cleaves the immobilized multi protein complexes. Binding is carried out again on a second column that consists of calmodulin beads. Subsequently, chelating calcium using EGTA elutes the native complex. Adapted from Huber LA, 2003 [2].

### *1.2.2.2 GST pull-down assay method*

The pull-down assay is a form of affinity purification method for studying physical protein interactions *in vitro* [79-81]. Pull-down assays are used as initial screening in identifying PPIs and also for affirming PPIs determined by other methods such as Co-IP. In this technique, a "bait" protein is used as a tag and captured on immobilized affinity ligand specific for the bait protein. This provides a secondary affinity support for purifying bait protein interactions. Incubation of the immobilized bait with putative "prey" proteins in a cell lysate is then carried out followed by protein elution with reducing buffers.

In the GST pull-down assay, glutathione S-transferase (GST) consisting of 220 amino acids [82], is used as a tag for studying interactions between proteins. GST DNA coding sequence is inserted next the gene coding for the protein of interest (at the N-terminus) and expressed in cells such as *Escherichia coli*. A fusion protein of the GST with protein of interest is thus formed after translation and transcription. The GST has a high affinity for the reduced form of glutathione, GSH. Therefore, the GST fusion proteins are purified by running a cell extract through a matrix of glutathione-coated beads, enabling the GST proteins to bind to beads and thereby isolating them from the rest of proteins in solution. Subsequently to purify the protein of interest, the GST fusion protein is washed and eluted with free GSH.

Pull-down assay aids in analysing strong or stable interactions in a variety of platforms, and also those that lack specific antibodies for protein complex immunoprecipitation (Co-IP). However, the GST pull-down has some disadvantages. The GST is relatively large in size (26 kDa) and when fused to the protein of interest, altering its native state [83, 84]. The bait proteins may have non-specific interact with other proteins and since pull-down assays are performed *in vitro* further investigations need to be performed to confirm these interactions *in vivo* [85].

#### *1.2.2.3 Protein complex immunoprecipitation (Co-IP) method*

Protein complex immunoprecipitation (Co-IP) is a technique widely used to detect, purify and analyse PPIs [62] using the principle of antibody-antigen reaction. In this technique, the protein complex is precipitated out as protein antigen (along with other macromolecules interacting in native state) using an antibody which specifically binds to that particular protein ("bait" protein) in a sample such as cell lysate [85]. Co-IP is used to target the bait protein with a specific antibody and thereby pull-out the entire protein complex to identify unknown interactions ("prey" proteins) in the large complex. This technique of pulling out all multiple proteins bound to the protein of interest is also known as "pull-down" technique. Co- IP technique is considered highly specific, relatively simple and compatible with most methods of downstream analysis while reagents can also be reused [86].

In this experiment, a specific antibody is added to the bait protein complex in a cell lysate and the complex is captured. Protein A or protein G covalently attached to beads is then used to immobilize the antibody. This is followed by washing of the beads with buffers to elute the bait and other proteins interacting with the bait protein. These bound proteins are commonly detected by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-

PAGE) and Western blot analysis [62, 87] as shown in Figure 1.3. MS is also used to identify the unknown proteins bound to bait protein [61].

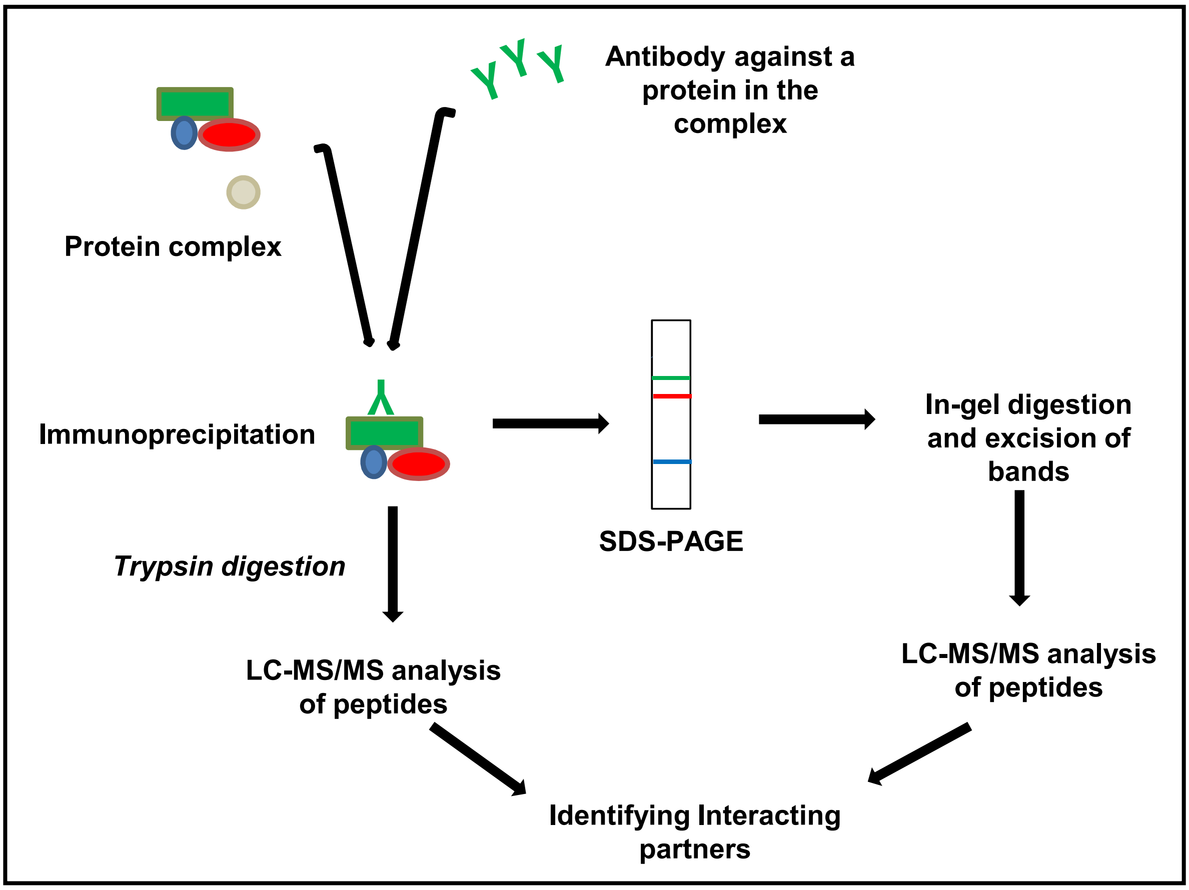

Antibody against a protein in the complex
Protein complex
Immunoprecipitation
SDS-PAGE
In-gel digestion and excision of bands
Trypsin digestion
LC-MS/MS analysis of peptides
LC-MS/MS analysis of peptides
Identifying Interacting partners

**Figure 1.3: Illustration of Co-IP technique for the identification of interacting proteins.** Adapted from Emri *et al*., 2011 [3].

Although, the Co-IP approach targets specific protein of interest using antibody with confidence in a physiological condition, it has some disadvantages. The signals for low-affinity PPIs may not be detected. Also in certain PPIs, there might be a third protein that may not necessarily be interacting. In addition, using specific antibodies for the predicted target proteins is highly important, which will otherwise lead to false interactions. Therefore, there may be occasional difficulties in obtaining antibodies of high specificity and avidity [86]. Moreover, the target protein needs to be accurately predicted. Furthermore, the Co-IP approach is not carried out in native membranes or in living cells [88].

### 1.2.3 Phage display

Phage display is a powerful technique which uses bacteriophages (viruses that infect bacteria) to link proteins with the genetic information that encodes them [89]. Figure 1.4 shows an example of phage display protocol. The phage display approach is similar to the two-hybrid system for high-throughput screening of protein interactions. This laboratory technique is used for studying protein-protein, protein-peptide and protein-DNA interactions. This technique also establishes binding preferences of peptide domains and is used to discover novel binding motifs [90]. Combinatorial peptide phage display uses highly diverse libraries to identify high-affinity ligands as potential inhibitors [91].

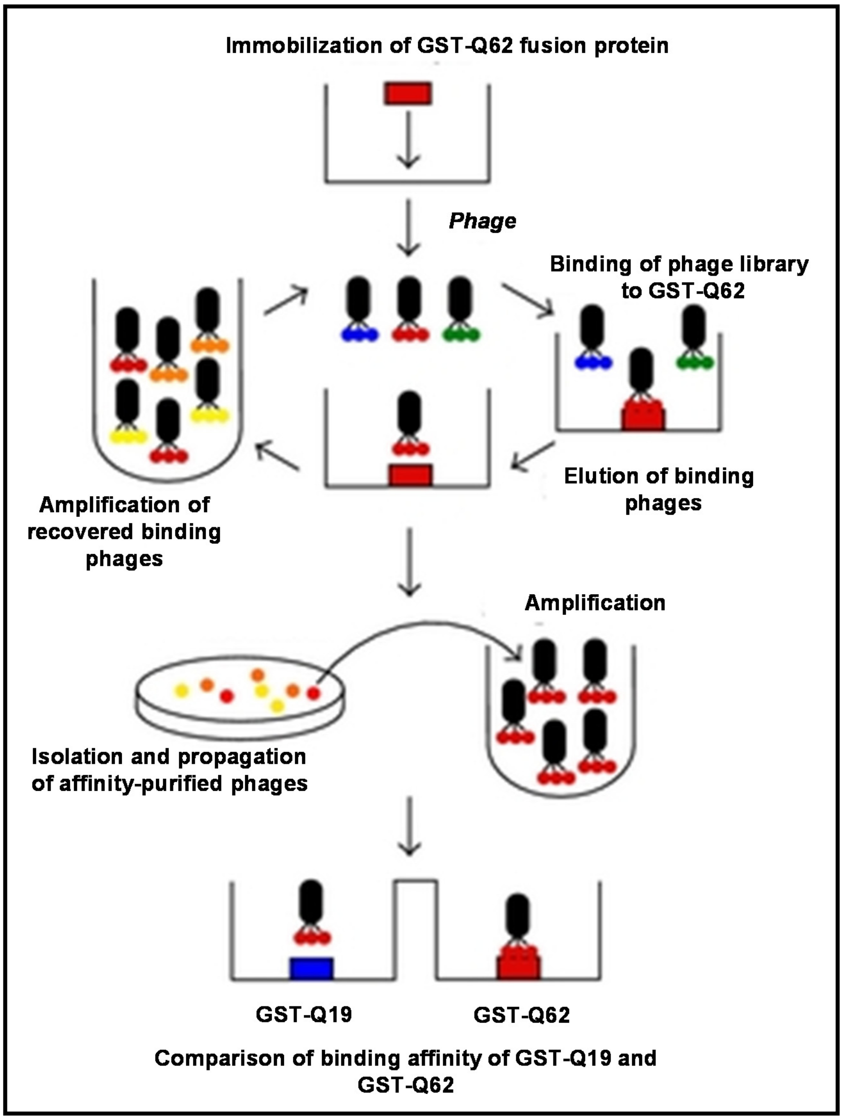

Immobilization of GST-Q62 fusion protein
Phage
Binding of phage library to GST-Q62
Elution of binding phages
Amplification of recovered binding phages
Amplification
Isolation and propagation of affinity-purified phages
GST-Q19
GST-Q62
Comparison of binding affinity of GST-Q19 and GST-Q62

**Figure 1.4: Illustration of phage display protocol for peptide identification.** Phage libraries expressing random AA are first screened for their binding to GST-Q62 through 4 rounds of binding, elution, and amplification. . Phage clones are then screened for their preferential binding to GST-Q62 compared to GST-Q19. Adapted from Popiel, 2011 [4].

In this technique, phage coat protein gene is inserted with a gene encoding the protein of interest, causing it to display the protein on its outside with the gene for the protein on its inside. This results in a link between the genotype and phenotype. The displaying phage is then screened for interactions with proteins, peptides or DNA sequences. Similarly, large libraries of proteins can be screened for interactions and amplified during *in vitro* selection. Bacteriophages such as M13 and fd filamentous phage [92, 93] are commonly used while T4 [94], T7 and λ phage have also been used.

In a proteomic phage display using the M13 phage [90], the input phage display libraries are constructed from cDNA, ORFs, or oligonucleotide arrays designed from a proteome of interest with peptides displayed on pVIII surface. Consequently, binding of phage occurs through interactions between displayed peptides and bait proteins. (4) the unbound phage are then washed, followed by elution of the bound phage through acidic or basic conditions or by adding actively growing host bacteria. Amplification of eluted phage and used for repeated cycles, to amplify specifically bound phage is then carried out. Subsequent NGS analysis of retained phage pools and/or Sanger sequencing of confirmed binders provides lists of binders from the target proteome.

Phage display is more of a survey tool, than an analytical one since it has some disadvantages as follows [95]. Predictions using combinatorial peptide phage display are not always accurate leading to time-consuming experimental validations of putative targets. The concentration of phage particles approaches $10^{12}$ per ml; however the molarity of displayed peptide is low (in Pico molar) leading to the need for using synthetic forms of peptides. The sequence complexity of the library is less than the number of recombinants in library and a negative result may be due to sparse sampling. The selection results can be inherently biased with biological selection against odd numbers of cysteine residues [96], runs of positive charges [97] and certain residues at fixed positions within the displayed peptide [98].

## 1.3 PROTEIN DATABASES AND RESOURCES

Large-scale genomics and proteomics studies has led to the determination of protein data through experimental and bioinformatics approaches and these have been systematically archived in several key databases (DBs) as discussed below.

### 1.3.1 General protein databases and resources

Several general databases contain useful information on proteins and their interactions ranging from published literature to protein sequences to X-ray and NMR structures of protein complexes. These general databases are discussed below.

#### *1.3.1.1 UniProt*

UniProt (www.uniprot.org), the *Universal Protein* Resource is a comprehensive, annotated and freely available database for high-quality information on protein sequence from genome sequencing projects and their biological functional information derived from literature. The UniProt knowledgebase (UniProtKB), a central repository is created by combining databases such as Swiss-Prot [99-101] developed by SIB (Swiss Institute of Bioinformatics) and EBI (European Bioinformatics Institute), TrEMBL (Translated EMBL Nucleotide Sequence Data Library- a computationally annotated supplement to Swiss-Prot) [99] and PIR-PSD (Protein Sequence Database produced by Protein Information Resource) [102-104].

The UniProt DB provides four core databases such as UniProtKB, UniParc, UniRef and UniMes.The UniProtKB is a protein database curated by UniProtKB/Swiss-Prot and UniProtKB/TrEMBL teams, providing protein sequences with consistent, accurate, rich sequence and functional annotation. The UniParc is a UniProt archive containing protein sequences from freely available protein sequence databases [105], and cross-referencing them to source databases to avoid redundancy during protein sequence retrieval. The UniRef, UniProt Reference Clusters, contains clustered sets of protein sequences from UniProtKB and UniParc records [106]. The UniMes, UniProt Metagenomic and Environmental Sequences DB contains metagenomic and environmental data [107]. As of April 2015 the UniProtKB/TrEMBL comprises 47452313 sequence entries, with 15721413695 amino acids.

#### *1.3.1.2 Swiss-Prot*

Swiss-Prot (http://www.expasy.ch/sprot/) is a manually curated protein sequence database with high-level of annotation [99]. The DB provides two classes of information, i.e. the core data (sequence data, citation information and taxonomic data) and annotation (the description of protein function, posttranslational modifications such as carbohydrates, phosphorylation, acetylation and GPI-anchor; domains and sites, e.g., calcium-binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains and kringle; secondary structure, e.g. α helix, β sheet; quaternary structure, i.e. homodimer, heterotrimer, etc.; similarities to other proteins; disease(s) associated with any number of deficiencies in the protein; sequence conflicts, variants, etc.). The source of information is publications reporting new sequence data and review articles to add annotations of families or groups of proteins, along with expert opinions or comments on specific groups of proteins. As of April

2015 UniProtKB/Swiss-Prot contains 548454 sequence entries, including 195409447 amino acids and are abstracted from 236665 published references.

#### *1.3.1.3 NeXtProt*

neXtProt (www.nextprot.org) is an on-line knowledge platform devoted to human proteins [108]. neXtProt provides high-quality information on all human proteins that are coded by 20,000 protein-coding genes in the human genome. This new resource contains a wealth of information such as biological function, subcellular location, expression, interactions and their role in diseases. The source of information at neXtProt is mainly from the UniProt Swiss-Prot database, while high-through studies especially proteomics also provide some data. As of May 2015, the neXtProt beta version contains 20,061 protein entries obtained from 439,129 published articles.

#### *1.3.1.4 PubMed*

PubMed (PubMed Central - PMC) is the central repository for biomedical literature from MEDLINE, life science journals, abstracts on biomedical topics and online books at the United States National Institutes of Health's National Library of Medicine (NIH/NLM). PubMed provides quality control on scientific publishing and only indexes/archives journals that meet PubMed's scientific standards. Each PubMed record is assigned a unique identifier (PMID). The resource comprises over 13.1 million records (listed with abstracts), and 14.2 million articles (having links to full-text) of which 3.8 million articles are freely-available with 24.6 million citations, as of May 2015.

#### *1.3.1.5 Protein Data Bank (PDB)*

PDB (www.rcsb.org/pdb) is the largest structural repository DB for three-dimensional (3D) structural data of large biological molecules, such as proteins and nucleic acids [109] [110]. It is an up-to-date archive for primary, secondary and tertiary structural data of biological macromolecules, obtained through X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world. A unique four letter code referred as PDB-ID or PDB-code is assigned to each structure deposited at the PDB. The PDB is a key resource for structural biology and structural genomics fields with hundreds of other DBs such as SCOP [111], CATH [112], SuperSite [113], PDB-ligand [114], PDBsum [115], GO [116] and ccPDB [117] categorising the primary information obtained at the PDB. A series of PDB related databases are reviewed elsewhere [118].

The deposition of structural complexes at the PDB is growing exponentially each year. As of May 2015, PDB contains 108395 biological structures. However, the data deposited at PDB contains few true or real complexes as compared to crystal artefacts despite significant development in determining the 3D structure of proteins [30]. Thus, obtaining a non-redundant dataset of protein complexes from the PDB is often a non-trivial task.

***PDB for PPI analysis:*** PDB provides information on the physical interactions between proteins. The structures deposited at the PDB contain experimental data to cross-validate their structural information. Therefore I have created an updated yet non-redundant dataset of known protein complexes from the PDB for our study.

### 1.3.2 PPI databases

PPIs are involved in several biological processes as in co-expression, gene regulation, metabolic pathways, cellular components and molecular co-evolution signifying a confounding data landscape. Hence, it is often necessary for the PPI databases to include new data, organize, curate and annotate the data for useful information, in order to link these data objects to biological reality. Some PPI databases also provide information such as functional gene links and domain level interactions [119]. Tuncbag and colleagues [119] reviewed PPI databases with their characteristics such as source organism, detection type, structural availability, interaction type and the number of interactions. Key PPI databases with information on their sources, interaction types and available database statistics are shown in Table 1.1. Brief summaries of the databases are provided in the following paragraphs.

The Database of Interacting Proteins (DIP) contains binary and multi-complex interaction information mined from experimentally techniques (Y2H, protein microarrays and TAP/MS) with nodes representing proteins and their interactions by edges [120]. DIP can also be accessed via Cytoscape plugin to view molecular interaction networks. Related databases include LiveDIP and Prolinks [121, 122]. The Biomolecular Interaction Network Database (BIND) Documents hand-curated molecular interactions data extracted from high-through experiments and gathered from literature [123, 124]. BIND records contain three classifications including interactions, complexes and pathways with over 206,859 unique interactions [125]. The Molecular INTeraction Database (MINT) [126, 127] database archives PPI reported from peer-reviewed articles. The interaction data is available as XML documents according to PSI-MI (Proteomics Standards Initiative-Molecular Interactions)

Level 1 and 2.5 standards, MITAB formatted files (a format defined by PSI-MI group with complexes represented as binary interactions) and as a simplified tab-delimited file.

**Table 1.1**: **List of currently available PPI databases**

| Database | Number of entries | URL |
|---|---|---|
| DIP | 27,390 proteins from 717 organisms and 78,735 interactions from 77,015 experiments and 7299 data sources (articles) (last update: Jan. 17, 2014) | http://dip.doe-mbi.ucla.edu |
| BIND | Over 206,859 unique interactions in PSI-MI BIND repository (last update: May 20, 2014) | http://bind.ca |
| MINT | 241,458 interactions  from 35,662 proteins spanning over 30 organisms (last update: Oct. 29, 2012) | http://mint.bio.uniroma2.it/mint |
| IntAct | 531,946 binary interactions from 13,807 curated publications and 1,298 biological complexes (last update: Aug. 25, 2015) | http://www.ebi.ac.uk/intact |
| BioGRID | 812,281 protein and genetic interactions, 27,034 chemical associations and 38,559 post translational modifications from major model organism species from 55,018 publications (last update: Sept. 1, 2015) | http://www.thebiogrid.org |
| HPRD | 41,327 PPIs with 30,047 protein entries, 93,710 PTMs, 112,158 protein expressions, 22,490 subcellular localization, 470 domains with 453,521 PubMed links (last update: April 13, 2010). | http://www.hprd.org |
| STRING | 9,643,763 proteins from 2031 organisms; 919,186,040 interactions (last update: April 12, 2015) | http://string-db.org |

IntAct [128] is a molecular interaction database providing experimentally determined PPIs across several species from literature curation or from direct user submissions. IntAct offers interaction data which complies with International Molecular Exchange Consortium (IMEx) guideline and the minimum information required to report a Molecular Interaction Experiment (MIMIx) standard. Biological General Repository for Interaction Datasets (BioGRID) [129] provides comprehensive curation of protein-protein and genetic interactions BioGRID release version 2.0 comprises >116,000 interactions from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. The Human Protein Reference Database (HPRD) [130] contains manually-curated protein interaction and pathways information of human proteins such as post-translation modification (PTM), subcellular localisation, expression, protein-domain architecture and disease in association with OMIM (Online Mendalian Inheritance in Man) database [131].

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [132] contains known and predicted PPIs using physical and functional associations obtained from genomic context, high-throughput experiments, (conserved) co-expression and previous knowledge (PubMed, mips etc.). The database contains experimental, predicted and transferred interactions, along with interactions obtained through text mining. The database also includes accessory information, such as protein domains and structures

Several other databases such as DOMINE (database of domain–peptide interactions, http://mint.bio.uniroma2.it/domino) [126], OPHID (Online predicted human interaction database, http://ophid.utoronto.ca) [133], HPID (Human Protein Interaction Database, http://www.hpid.org) [134], and PIPS (human protein-protein interaction database, http://www.compbio.dundee.ac.uk/www-pips) [135], provide information on the protein interaction data generated through high and low throughput experiments, compiled from literature or from computational predictions. The Munich Information Centre for Protein Sequences (MIPS: http://mips.gsf.de) archives genome and protein sequences [136]. Comprehensive Yeast Genome Database (CYGD) of MIPS consists of two separate sets of yeast protein interactions, one with manually curated interaction and the other set generated through high throughput experiments [137].

Protein complexes deposited at the PDB is most widely used for structural interaction analysis and functional characterisation of complexes. Databases such as BID (Binding Interface Database, http://tsailab.chem.pacific.edu/wikiBID/) [138] and ICBS (Inter-Chain Beta-Sheets, http://icbs.ics.uci.edu/) [139] provide structural information of protein interactions with reference to literature.

***Usage of these DBs:*** Based on the availability of data organized and presented in these databases, users can choose an appropriate database for specific data retrieval or to validate their predicted interactions. IntAct, BioGRID and STRING are currently regularly updated.

## 1.4 CURRENT TRENDS IN COMPUTATIONAL PPI ANALYSIS AND PREDICTION

The current trends in computational analysis of PPIs using structural and sequence information of existing PPI data, their deposition in various databases and their utilization in the understanding of binding principles of PPI is discoursed. Application of PPI analysis in the prediction of interaction partners and sites is also discussed. Formation of a PPI complex is shown in Figure 1.5.

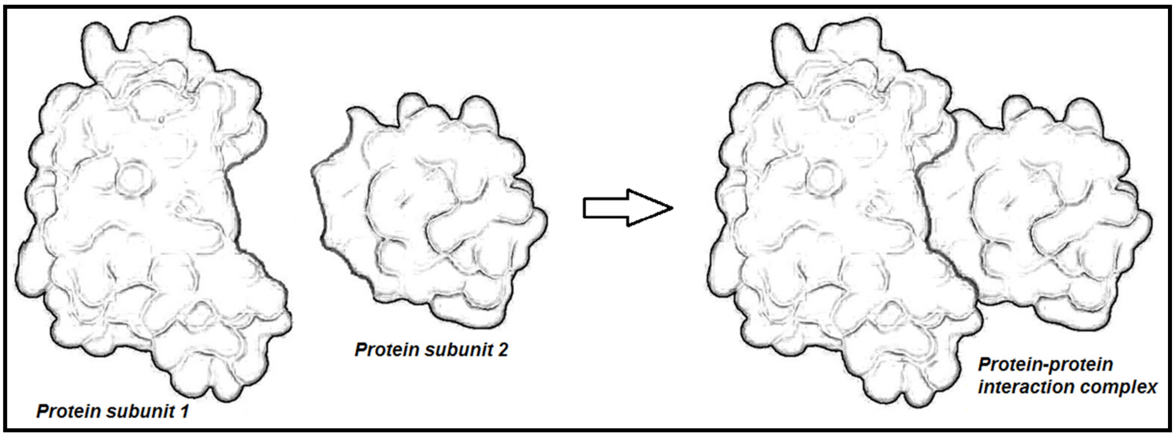


**Figure 1.5: PPI complex formation is represented.** The protein subunit 1 and subunit 2 come closer to form a PPI complex. Adapted from Sowmya *et al.*, 2010 [5].

### 1.4.1 Classification of PPI complexes

PPI complexes can be broadly classified into two main classes based on the size of protein subunits and their composition into homo-oligomers and hetero-oligomers. PPIs are also classified based on affinity and lifetime of their association into obligate or non-obligate and permanent or transient complexes, respectively [50], as shown in Figure 1.6 [6]. Moreover, aminoacid composition alone can discriminate the different classes of complexes (six types of PPI interactions) such as interactions within the same structural domain and between different domains, permanent and transient interfaces, homo- and hetero-obligomers (obligomer is a polymer consisting of monomeric units) [140]. Therefore, discrimination of the different types of protein complexes to identify structural diversity of PPIs for analysis and prediction studies helps gain insights into their nature of binding and their role in a biological cell [30, 50, 141].

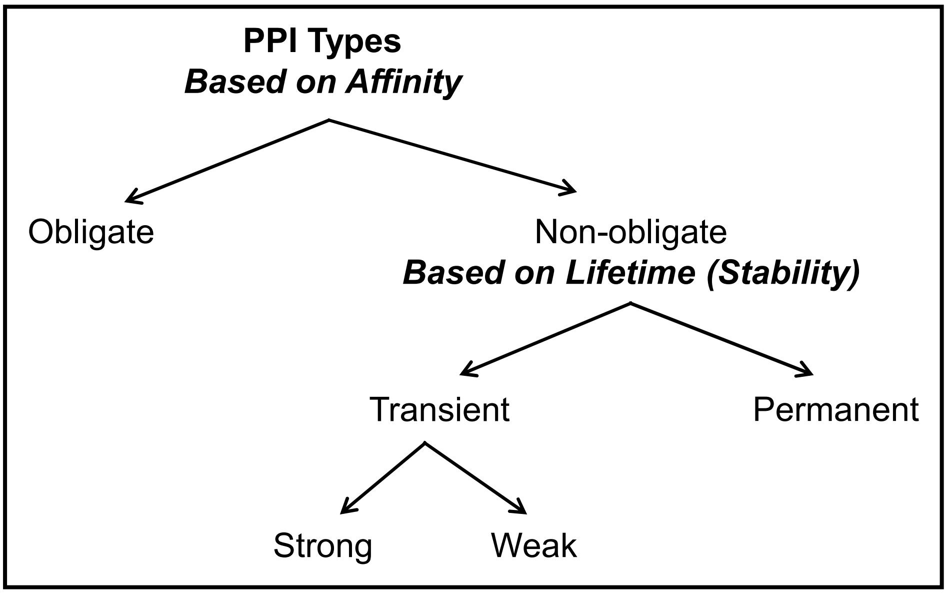


**Figure 1.6: Representation of the diverse types of PPI based on affinity and stability.** Adapted from Ozbabacan *et al*., 2011 [6].

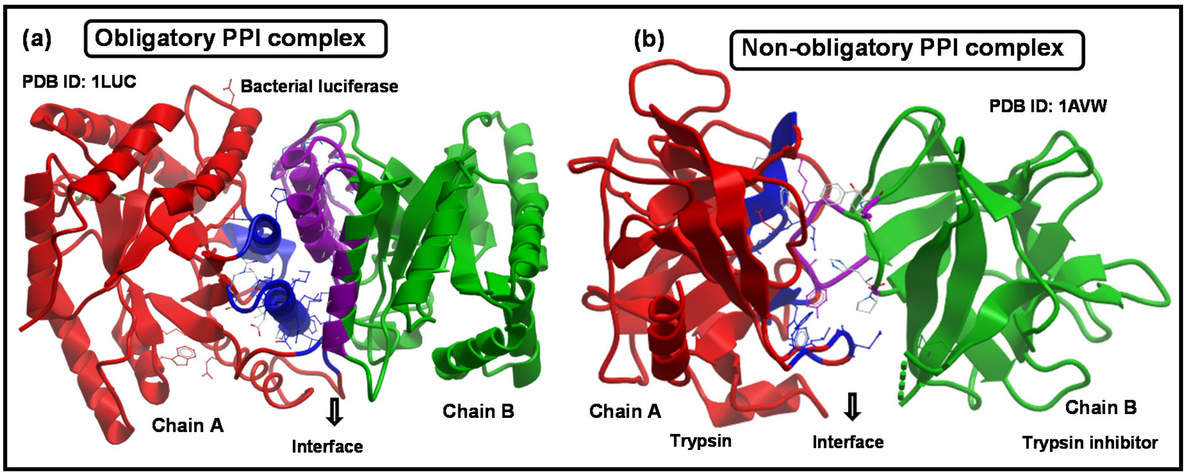


**Figure 1.7: Examples of obligate PPI complex and non-obligate PPI complex is shown in cartoon representation.** a) An obligate interface of bacterial luciferase with chain A and chain B (coloured in blue and magenta, respectively) interacting to form a stable interface to catalyse the oxidation of long-chain aldehydes. b) A non-obligate complex of pancreatic trypsin/trypsin inhibitor interacting transiently at the interface coloured as in (a).

#### *1.4.1.1 Obligate and non-obligate complexes*

PPIs are classified based on their affinity into obligate or non-obligate complexes [50] (examples of obligate and non-obligate protein dimers are shown in Figure 1.7). Obligate PPI complexes consist of protein subunits (or monomers) that do not exist as independent monomers or stable structures *in vivo*. However, non-obligate PPI complexes constitute protein subunits that can exist as independent structural monomers. For example, Krishna and Aravind showed how Ku protein is likely to bind to DNA to form an obligate dimer [142]. Examples of non-obligate complexes include signalling proteins, where constituent proteins dissociate into individual protein monomers after propagating signals [6].

#### *1.4.1.2 Permanent and transient complexes*

PPIs are also classified into permanent and transient complexes based on their lifetime of association [6, 50]. Permanent complexes usually exist in stable and irreversible forms, whereas, transient complexes associate and dissociate temporarily *in vivo* [6]. Transient PPI complexes can be further classified into strong and weak based on their affinities and lifetime of association [50]. Discrimination of PPI into permanent and transient complexes can be done based on structural features such as interface geometrical and physicochemical properties and also sequence properties such as amino acid substitutions [23, 50, 143, 144]. These properties are the changes in accessible surface area (ΔASA) and planarity, size and shape properties, gap volume index, polarity [23, 50, 143], hydrophobicity, average number of hydrogen bonds, the number of discontinuous segments in the interface, secondary structures and the extent of conformational change upon binding [6].

Permanent complexes are usually obligate complexes with stable interfaces and hence interchangeably used in literature. Similarly, non-obligate complexes are abundantly transient in nature with few examples of permanent complexes [22, 50]. Ozbabacan and colleagues [6] comprehensively reviewed transient PPI complexes with examples of transient and permanent complexes. Example of permanent (Ascaris pepsin inhibitor-3 bound to porcine pepsin protein complex) and transient complex (An ARF1-GDP bound to Sec7 domain complexed with brefeldin A) is shown in Figure 1.8.

#### *1.4.1.3 Homo-oligomers and Hetero-oligomers*

Identical protein subunits (or chains) interact to form homo-oligomer complexes. The folding and binding mechanism of homo-oligomers is different from hetero-oligomer complexes. The homodimers (association of two identical subunits, Figure 1.9a) are formed through three folding mechanisms such as 2S (2S without any intermediate), 3SMI (3-state with monomer intermediate) and 3SDI (3-state with dimer intermediate) [145]. Hence, analysis of the homo-oligomeric association and folding mechanism in known structural complexes, helps predict folding mechanism given structural complexes [146]. The binding modes of homodimers play a key role in distinguishing them from heteromeric complexes based on electrostatics [147], as heterodimers, composed of different monomers, can carry opposite net charge, while homodimers, arising from two identical monomers cannot.

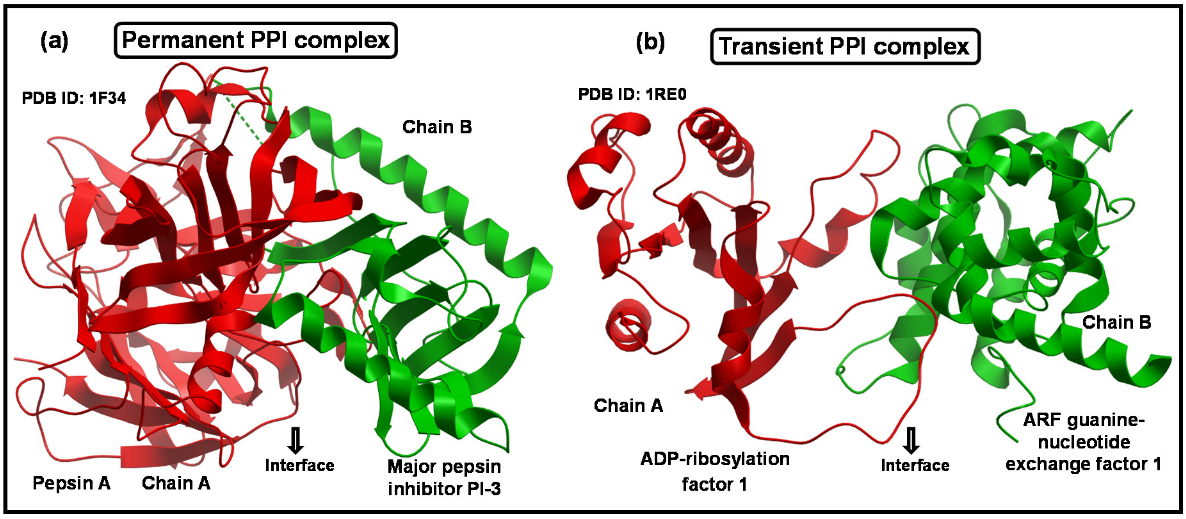


**Figure 1.8: Examples of permanent PPI complex and transient PPI complex is shown in cartoon representation.** a) A permanent complex of Ascaris pepsin inhibitor-3 bound to porcine pepsin. b) A transient complex of ARF1-GDP bound to Sec7 domain complexed with brefeldin A.

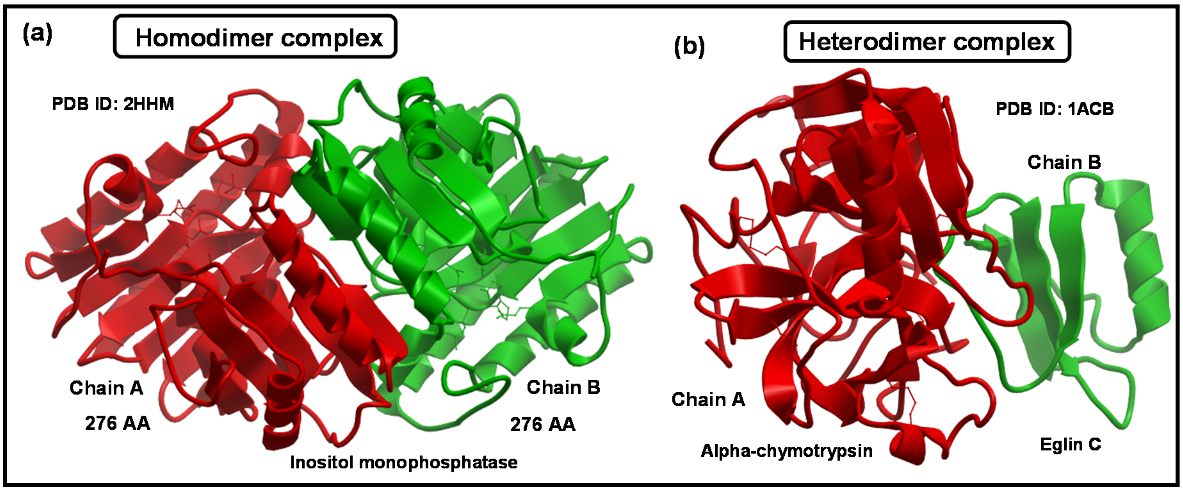


**Figure 1.9: Examples of homodimer and heterodimer complexes are shown in cartoon representation.** a) A homodimer protein complex (inositol monophosphatase) is formed from the association of chain A (red) and chain B (green) of identical size (276 amino acids (AA) each). b) A heterodimer complex is formed from the association of chain A (α-chymotrypsin) and chain B (Eglin C) coloured in red and green respectively.

Non-identical protein subunits (or chains) non-covalently interact to form hetero-oligomers (shown in Figure 1.9) with diverse functionality. The stability of heteromeric complexes may vary, and these proteins constitute a single macromolecular assembly [6].

#### *1.4.1.3.1 Heterodimers and their advantages in PPI analysis*

Heterodimeric interactions are commonly found in enzyme-inhibitors, enzyme complexes, antibody-antigen, signal proteins and cell cycle proteins [23, 39]. Heterodimer complexes are more intriguing than homodimers since they also include transient complexes [30]. Therefore, heterodimers are often used in studying PPIs for applications in PPI prediction [146].

#### *1.4.1.3.2 Dimers over multimers for PPI analysis*

Multimeric proteins represent different levels of interaction in a living cell. Also, multimeric interactions are, in general, weaker with temporary contacts among the interacting protein subunits, except in the case of very large complexes (e.g. the proteasome) or viral surfaces. However, dimeric interactions are amongst the strongest and most extensive in nature [23, 32], suggesting that dimeric proteins are long-lived, with isolated oligomer subunits rarely achieving their biological function in monomer state.

Hence, I have created an updated non-redundant dataset of heterodimers from the PDB and analysed them to gain insights into PPIs (detailed in Chapter 3 and Chapter 4).

### 1.4.2 Datasets for PPI analysis and prediction

PPIs are extensively studied using non-redundant datasets of structural complexes from protein structural repository databases such as PDB [148]. Also, PDB contains structural data that are frequently unorganized despite significant efforts to structure and organize the data. Therefore, it is difficult to maintain a default procedure to mine for a reliable dataset or in using a standard dataset for PPI analysis and to train predictors. Hence, creating an updated yet non-redundant dataset representing protein multimers from PDB is a non-trivial task and an important step in PPI studies [24]. The different levels of structural data available at the PDB are shown in Figure 1.10.

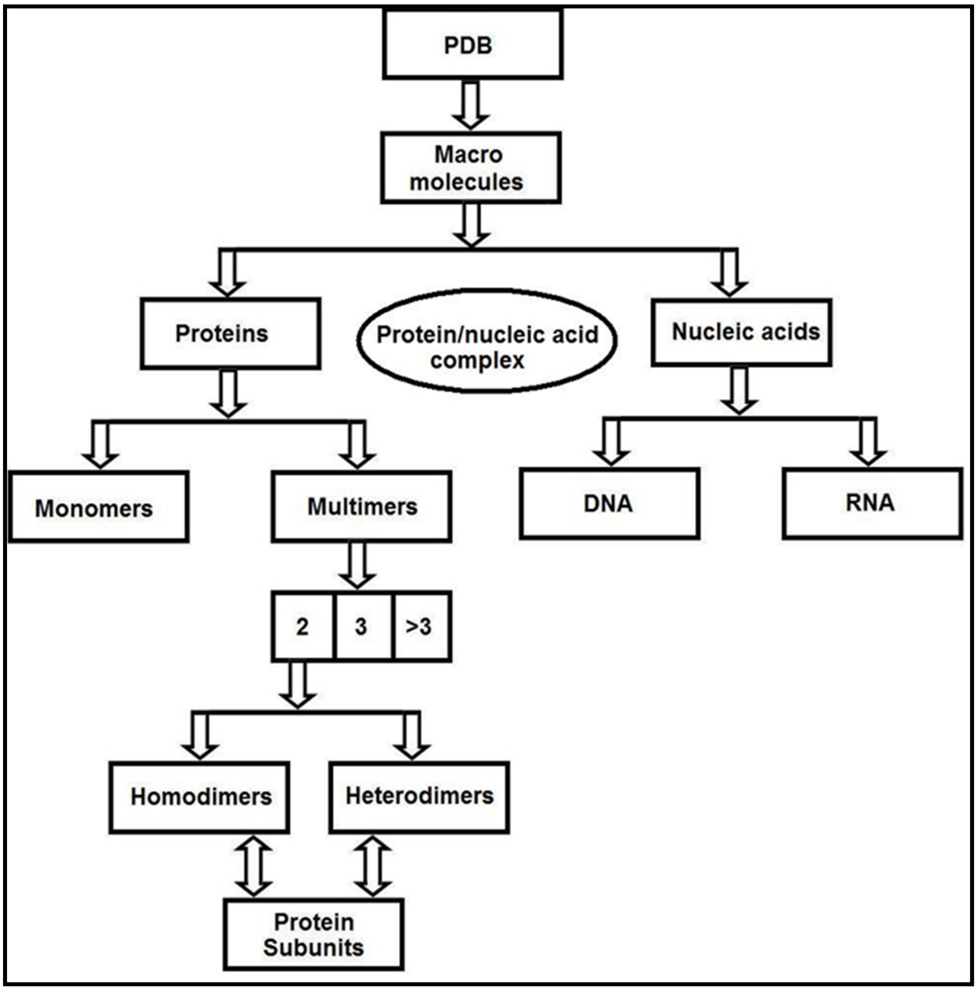


**Figure 1.10: Different levels of structural complexes available at PDB.**

The datasets collected/created by different groups for PPI studies contains heterogeneous data (disproportionate mixture of homodimers and heterodimers). The classical work by Chothia and Janin in 1975 [31] with a modest dataset of three protein complexes, has paved the path towards the understanding of PPI using known structural complexes. Jones and Thornton (1996) [23] used a dataset of 59 protein complexes consisting of 32 homodimers, 10 enzyme-inhibitor, 6 antibody-protein complexes and 11 hetero complexes to study the difference in various structural features among the classes. Xu and colleagues (1997) [38] generated a dataset of 319 protein-protein interfaces to analyse the features such as hydrogen bonds and salt bridges to gain insight into the specificity of protein-protein associations. In parallel, Tsai and colleagues (1997) [149] from the same group have used a dataset of 362 protein-protein interfaces to perform a statistical analysis of hydrophobic effect at the interfaces. Lo Conte and colleagues [39] used 75 protein-protein complexes to study interface size distribution.

Similarly, various other groups have assembled datasets of X-ray crystal structures from the PDB to examine the different properties of subunit-subunit interactions [24, 43, 143, 150-156]. Protein-protein docking benchmark datasets [157] have also been widely used to understand the binding principles of PPI [48]. Grouping of dataset to study bias in interface properties in the different types of PPI has also been carried out [40, 50, 57]. Table 1.2 shows the heterogeneous datasets created/collected over the decades by various groups for PPI analysis and prediction (in chronological order).

**Table 1.2: PPI datasets created/collected by various groups**

| Authors | Year | Dataset | Reference |
|---|---|---|---|
| Chothia and Janin | 1975 | 3 (insulin dimer, trypsin-PTI, α/β oxyhaemoglobin) | [31] |
| Chothia *et al.* | 1976 | 2 (horse methemoglobin, human hemoglobin) | [34] |
| Miller *et al.* | 1987 | 11 dimer, 9 tetramers, 2 hexamer, 1 octamer | [35] |
| Janin and Chothia | 1990 | 15 protease inhibitors, 4 antigen-antibody complexes | [36] |
| Jones and Thornton | 1995 | 32 complexes | [32] |
| Janin and Rodier | 1995 | 152 crystal forms of monomeric protein | [158] |
| Jones and Thornton | 1996 | 59 complexes (32 homodimers; 10 enzyme – inhibitor; 6 antibody – protein complexes; 11 hetero complexes) | [23] |
| Xu *et al.* | 1997 | 319 protein-protein interfaces | [38] |
| Tsai *et al.* | 1997 | 362 non-redundant protein-protein interfaces, 57 symmetry-related oligomeric interfaces | [149] |
| Dasgupta *et al.* | 1997 | 58 oligomeric proteins, 223 protein crystal structures | [159] |
| Linzaad and Argos | 1997 | 59 protein complexes with 159 polypeptide chains | [51] |

| Authors | Year | Dataset | Reference |
|---|---|---|---|
| Jones and Thornton | 1997 | Protomers from 28 homodimers, large and small protomers from 11 and 14 hetero-complexes respectively, and antigens from 6 antibody-antigen complexes. | [160] |
| Lo Conte *et al.* | 1999 | 75 complexes (24 protease inhibitors, 19 antigen-antibody) 32 others (9 enzyme inhibitors, 11 signal transduction) | [39] |
| Valdar and Thornton | 2001 | 53 families of homodimers and 65 families of monomers | [161] |
| Zhou and Shan | 2001 | 615 pairs of non-homologous complex-forming proteins | [162] |
| Fariselli *et al.* | 2002 | 226 heterodimers | [163] |
| Chakrabarti and Janin | 2002 | 70 protein-protein complexes | [150] |
| Brinda *et al.* | 2002 | 20 homodimers | [151] |
| Bahadur RP *et al.* | 2003 | 122 homodimers | [41] |
| Mintseris *et al.* | 2003 | 59 test cases: 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and 7 difficult test cases. 31 of the test cases (with receptor and ligand) are classified as 16 enzyme-inhibitor, 5 antibody-antigen, 5 others difficult | [164] |
| Mintseris &Weng | 2003 | 209 protein complexes | [165] |
| Nooren and Thornton | 2003 | 39 (16 experimentally validated "weak" transient homodimers and 23 functionally validated transient heterodimers) | [143] |
| Caffrey *et al.* | 2004 | 64 complexes (42 homodimers, 12 heterodimers, 10 transient complexes) | [153] |

| Authors | Year | Dataset | Reference |
|---|---|---|---|
| Bahadur *et al.* | 2004 | 70 protein-protein complexes and 122 homodimers | [152] |
| Zhanhua et al. | 2005 | 65 heterodimers | [166] |
| Zhanhua *et al.* | 2005 | 156 heterodimers, 170 homodimers | [43] |
| Mintseris *et al.* | 2005 | 72 unbound-unbound cases, with 52 rigid-body cases, 13 medium-difficulty cases, and 7 high-difficulty cases with substantial conformational change, and 12 antibody-antigen test cases | [167] |
| Zhu *et al.* | 2006 | Dataset of 243 protein interactions | [168] |
| Pal *et al.* | 2007 | 204 protein complexes | [44] |
| Li *et al.* | 2007 | 1276 non-redundant hetero-complex protein chains | [169] |
| Keskin *et al.* | 2008 | 3799 structurally non-redundant interfaces | [170] |
| Hwang *et al.* | 2008 | 124 unbound-unbound test cases classified into 88 rigid-body cases, 19 medium-difficulty cases, and 17 difficult cases | [154] |
| Choi *et al.* | 2009 | 2646 protein interfaces with the classification of homodimeric/heterodimeric and obligatory/transient interactions | [171] |
| Reynolds *et al.* | 2009 | 220 heterodimers, 534 homodimers | [155] |
| Gromiha *et al.* | 2009 | 153 heterodimers | [172] |
| Liu *et al.* | 2010 | 130 protein chains from transient complexes | [173] |
| Guharoy and Chakrabarti | 2010 | 122 homodimer, 204 heterodimer | [156] |
| Hwang *et al.* | 2010 | 176 (52 new complexes added to the 124 cases of Benchmark 3.0) | [157] |
| Kastritis *et al.* | 2011 | 144 protein-protein complexes | [174] |

| Authors | Year | Dataset | Reference |
|---|---|---|---|
| Sowmya *et al.* | 2011 | 192 heterodimer complexes | [24] |
| Swapna *et al.* | 2012 | 223 homodimers | [175] |
| Swapna *et al.* | 2012 | 76 protein-protein complexes | [176] |
| Chen *et al.* | 2013 | 113 heterodimer complexes | [49] |
| Gromiha *et al.* | 2014 | 185 protein–protein complexes | [177] |
| Du *et al.* | 2015 | 270 hetero complexes and 289 homo complexes | [178] |

### 1.4.3 Discerning features of PPI interfaces

Protein associates with another protein forming a PPI complex to achieve a specific biological function. A stable interface is often formed between the two interacting protein subunits. Figure 1.11 shows an example of a typical PPI heterodimer complex. The ability of a protein to interact with its partner depends on various physical and chemical features. These factors determining the formation of the interface is often multi-parametric in nature.

The physicochemical properties governing subunit-subunit interactions in a protein complex have been extensively studied over the past few decades and described in a number of studies elsewhere [31, 32, 35, 36, 39, 149, 179]. Protein-protein interfaces have several distinct features that distinguish it from the rest of the protein surface. The interface features have been considered to be additive in nature and therefore, these combinatorial features are often investigated to determine the major driving forces for binding and also towards using these features to train predictors for futuristic PPI predictions.

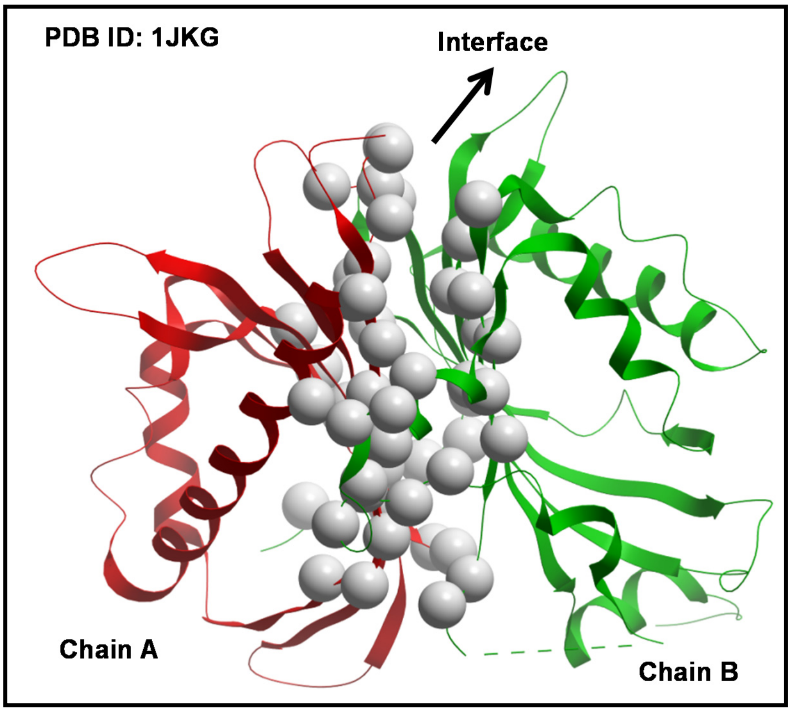


**Figure 1.11: A pictorial representation of a protein-protein interaction complex.** Rat anionic trypsin heterodimer (PDB ID: 1JKG) is shown here. The interacting residues at the protein interface are shown in space-filling or CPK representation with interface residues as white balls. The protein chains are shown in cartoon representation with chain A and chain B coloured in red and green respectively.

### 1.4.3.1 Structural features of PPI

The physical and chemical features that contribute to formation of an interface include shape, shape complementarity, planarity, sphericity, interface size, interface area, gap volume, gap index, residue side-chain packing, residue propensities (frequency of amino acid residue type), electrostatics, hydrogen bonds, salt bridges and disulphide linkages. These structural interface features have been studied extensively by various groups using datasets primarily from the PDB.

Figure 1.12 shows the protein-protein interface, formed between two proteins, labelled Subunit 1 and Subunit 2.

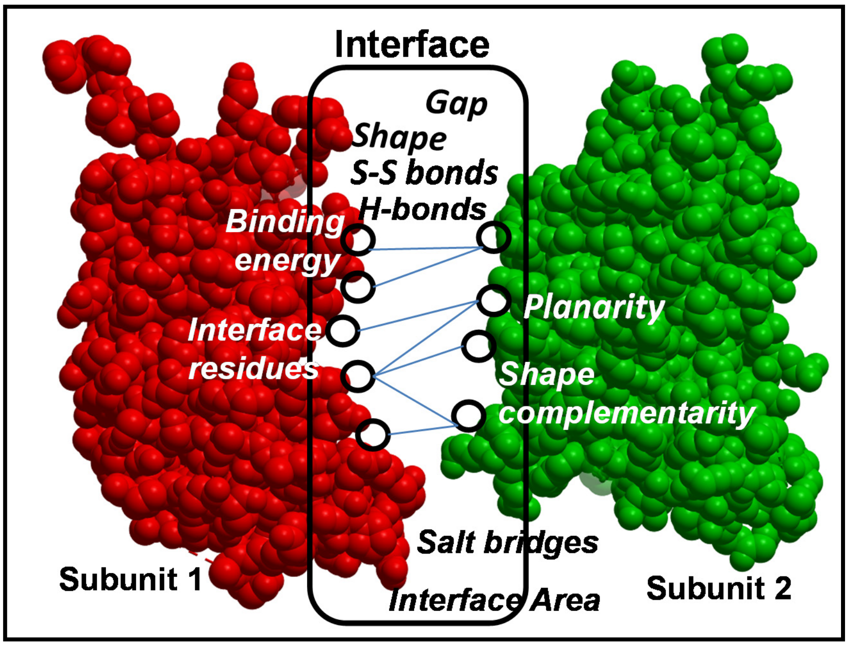


**Figure 1.12: A description of the interface in a PPI complex is shown.** Subunits 1 and 2 interact at specific binding sites to form a stable interface, involving interface residues. These interface residues may be hydrophobic, hydrophilic or amphipathic in nature. These interface residues form hydrogen bonds, salt bridges and/or disulphide linkages [7].

#### *1.4.3.1.1 Interface Area*

A stable interface is formed between two interacting subunits in a PPI complex. The interface is identified by calculating the change in solvent accessible surface area (ΔASA) upon complex formation (Figure 1.13). A probe radius of 1.4 Å is used to calculate ASA with Lee and Richards implementation [180].

The ΔASA of a protein complex is calculated as follows:

$$\Delta\text{ASA (complex)} = \frac{[\text{ASA (subunit 1)} + \text{ASA (subunit 2)} - \text{ASA (complex)}]}{2} \qquad (1)$$

The solvent accessible surface area (ASA) of subunit 1 and subunit 2 are calculated, followed by the ASA of the complex. Thus, the ΔASA (complex) or the interface area (B/2) of the complex is the surface that becomes buried when two proteins interact.

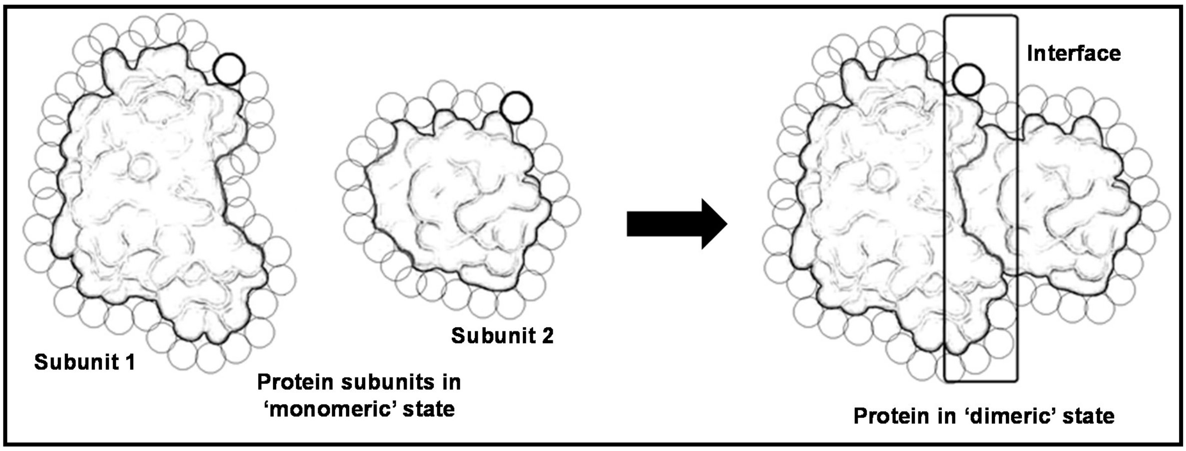

Interface
Subunit 2
Subunit 1
Protein subunits in 'monomeric' state
Protein in 'dimeric' state

**Figure 1.13: Illustration of the delta ASA (ΔASA) analysis.** The Accessible Surface Area (ASA) is calculated individually for subunit 1, subunit 2 and protein complex. Adapted from Sowmya *et al.,* 2010 [5].

Jones and Thornton showed that the average size of a homodimeric complex ranges from 368 – 4746 Å$^2$, while heterodimers have an interface area within 639 – 3228 Å$^2$ [23]. The interface area is also shown to range between 670 – 5540 Å$^2$ based on 23 oligomeric protein complexes [159]. The interface area in heterodimeric protein complexes are in a range of 1200 – 2000 Å$^2$, denoted as the standard size of an interface [39]. Bahadur and colleagues showed that the range of interface area extends from 500 – 7000 Å$^2$, with a mean of 1970 Å$^2$ [34]. Caffrey and colleagues showed that interface area ranges from 415 – 2361 Å$^2$ for heterodimer complexes, 550 – 4718 Å$^2$ for homodimers and 423 – 2361 Å$^2$ for transient complexes [153]. These different values have been reported at different time points and based on different datasets of complexes.

In general, the interfaces in homodimeric complexes are on average 2-fold larger than heterodimeric protein complexes and about 2.5-fold larger than the crystal-packing interfaces of monomeric proteins. Therefore, based on the 2008 review of Bahadur and Zacharias, the size of interfaces or interface area ranges from as small as 800 Å$^2$ to very large interface area of more than 10,000 Å$^2$ (in a few homodimeric complexes) [181]. Stronger protein subunit binding was commonly associated with larger interface areas. Also, homodimers are known to have larger interfaces than heterodimers [37, 43]. Thus, interfaces can vary based on different datasets and their features such as resolution, size and type of dataset [182]. The accessible and buried surface area is related to molecular weight [35]. Also, stronger protein-protein binding was associated with larger interface areas [32].

#### *1.4.3.1.2 Interface residues*

The interface is composed of interacting residues from either subunit/chain known as interface residues. The interface residue becomes inaccessible to the solvent upon protein-protein binding. Thus, the interface residues are filtered based on the criteria that their ΔASA > 0.1Å$^2$ [150]. Porollo and Meller [55] filtered interface residues based on the criteria that their relative ΔASA is at least 4% and not less than 5 Å$^2$ upon complex formation. The interface residues may be hydrophobic, hydrophilic or amphipathic in nature (with equal distributions of hydrophobic and hydrophilic residues) [5]. The average number of interface residues is 44.4 in homodimers and 42.2 in heterodimer complexes [182]. The number of residues forming the interface (or interface residues) is proportional to interface area [41, 150, 151].

#### *1.4.3.1.3 Interface patches*

Interfaces patches are widely used to characterise protein subunit interfaces and also in PPI binding site predictions. Jones and Thornton (1997) studied PPI sites using surface patch analysis to compare how interaction patches differed from the rest of the surface [37]. Hydrophobic patches (clusters of hydrophobic residues deemed accessible on a given protein surface) on protein subunit interfaces were known to be involved on multimeric interfaces in 90% of complexes [51]. Chakrabarti and Janin dissected protein recognition sites and described recognition patches in protein interfaces. These recognition sites are refined for a typical interaction 'patch' as having an area of at least 800 Å$^2$ involving more than 20 residues and less than 100 atoms [150], from protein chains of at least 50 residues.. The interface patches are composed of a core and a rim. The core residues are shown to have a distant composition than the rest of the surface, while the rim isolates the patch from the solvent. An interface patch is composed of 47±11 residues or 23 residues per recognition site. Dominant peptide segments were observed at the interfaces of homodimer and crystal contacts [44]. Guharoy and Chakrabarti (2010) also documented that conserved residues are not randomly distributed over the whole interface but form distinct clusters using a dataset of 122 homodimer and 204 heterodimer complexes [156].

#### *1.4.3.1.4 Hydrophobic/non-polar/apolar interfaces*

The interface residue composition determines the chemical nature of the interface. Hence, the nature of residues at the interface determines the hydrophobic effect of protein association. The classical work by Chothia and Janin (1975) showed that protein interact by burying large amounts of hydrophobic surface areas [31]. Chothia and colleagues also demonstrated the role of hydrophobic interface to allosteric mechanism in human deoxy-haemoglobin [34]. Similarly, Jones and Thornton [32] reaffirmed the fact that hydrophobic residues are more at interface than surface but less than interior, using a dataset of 32 complexes. They also assumed that proteins associate with each other through hydrophobic patches[23]. Tsai *et al.* (1997) also recognized the role of interface hydrophobic residues (although not as much as in protein folding) in binding, and found exceptions where there is no sign of significant hydrophobic contribution at the interface [149]. Furthermore, large and strong interface hydrophobic patches (cluster of neighbouring apolar atoms at the interface) have been shown to be dominating feature at the protein interfaces [51].

Compared to heterodimers, homodimers are known to have larger interfaces with predominant hydrophobic residues at their interfaces [37, 43]. Thus, the hydrophobic effect is known to play a major role in the formation of PPI complexes. The hydrophobic distribution patterns for interfaces is quite different with some demonstrating a single large patch of hydrophobic residues surrounding by polar residues, while some others show 1-3 patches of hydrophobic residue distributions, linked to water molecules and hydrogen bonds across the interface [182]. Most studies analysed the average hydrophobicity over a diverse set of PPI complexes, blurring the information on the contributions of hydrophobic effect to individual proteins complex stability and interactions.

#### *1.4.3.1.5 Hydrophilic/polar interfaces*

The interface residues in some PPI complexes are abundantly hydrophilic/polar, with predominant polar interactions at the interface. Lo Conte and colleagues (1999) showed that protein-protein interfaces are significantly polar as well as non-polar, with characteristics similar to the protein surface [39]. The fractional distribution of hydrophobic/non-polar, hydrophilic/polar and charged residues in homodimer and heterodimer interfaces was also demonstrated [43]. Hydrophilic/polar residues (W, C, H, Q, N, Y, S), except for T, were observed to occur dominantly in heterodimer. Moreover, heterodimer complexes were observed to commonly associate by polar interactions and not by hydrophobic interactions as in homodimers [5, 43].

#### *1.4.3.1.6 Hydrogen bonds (H-bonds)*

Intermolecular hydrogen bonds (H-bonds) formed between the two interacting subunits/chains play an essential role in determining stability at the interface. The hydrogen atoms covalently bound between two electronegative atoms from two different chains and contributing to electrostatics are within a distance of 4 Å. The covalent bond is energetically favourable as it includes polarization energy, covalent energy and particularly the electrostatic energy [182]. The energy of an average inter-molecular h-bond is generally small, 20 KJ/mol (5 kcal/mol) [32], however play an important role in PPIs. The number of hydrogen bonds identified varies in different studies.

Janin and Chothia showed that interfaces have 8 – 13 h-bonds with an average of 10 hydrogen binds per complex [36]. Jones and Thornton showed that interfaces have 0 – 46 h-bonds with an average of 0.88 per 100 Å$^2$ of interface area, with a Pearson correlation coefficient (*r*) value of 0.77 between h-bonds and interface area [32]. Xu and colleagues (1997) also showed 11 H-bonds per subunit, with r value of 0.89 between H-bonds and interface area [38]. On an average, 10.1 H-bonds are formed at a protein-protein interface, with one H-bond per 170 Å$^2$ interface area with an *r* value of 0.84 observed between H-bonds and interface area [39]. The r value between H-bonds and interface area calculated using different dataset size and nature of data varies from 0.75 to 0.89 [23, 32, 38, 39, 43, 166], with an average of 0.24 H-bonds per interface residue in heterodimers. High H-bond density per interface residue (0.64) with dominant charged and hydrophilic/polar residues at the heterodimer protein interfaces is also demonstrated [43]. Therefore, documentation of inter-molecular hydrogen bonds among different protein complexes helps determine stability and evaluate predictions [23, 32, 38, 39, 41].

#### *1.4.3.1.7 Salt bridges*

Intermolecular salt bridges formed between the two interacting subunits/chains also contribute to the stability and electrostatics of protein complex. The salt bridges are formed between two oppositely charged side-chain atoms (i.e., basic and acidic amino acids) within a distance of 4Å. An average of 2.0 salt bridges per interface has been documented by Xu and colleagues [38]. Salt bridges are also known to provide favourable free energy to binding, however an isolated charge buried in a protein interface could considerably destabilize binding due to the desolvation effect [38]. Thus salt bridges and hydrogen bonds contribute to high selectivity and specificity in protein binding.

#### *1.4.3.1.8 Shape complementarity*

Shape complementarity and geometric complementarity between interacting proteins has been observed to have a very important effect on PPIs for several decades [23, 183-185]. Shape complementarity has been characterised by the size of interface area, buried water molecules and packing density of interface atoms [183, 186, 187]. Sternberg and Gabb scored potential complexes on the basis of shape complementarity and favourable electrostatic interactions using Fourier correlation theory [188]. Li and colleagues [185] demonstrated the role of shape complementarity in protein complexes They showed that when two membrane proteins with shape complementarity come close to each other, the lipid chains in the membrane between the two chains would have restricted conformation, therefore, tend to leave the gap between the proteins to maximize configuration entropy. This yields an effective entropy-induced PPI enhancing protein binding.

#### *1.4.3.1.9 Gap volume and gap index*

The complementarity of interacting protein surfaces can also calculated using gap index [23]. The volume enclosed between the two interacting subunits is called the gap volume. The gap volume is generally calculated using SURFNET program with a procedure developed by Laskowski [189]. This algorithm runs a series of spheres (of maximum radius 5Å) between the surfaces of each interacting protein subunit atoms, such that the surface is contact with the surfaces of the atoms on either side. The interception of other atoms causes the size of sphere to reduce accordingly and is discarded if it falls below a minimum radius (1Å). Thus, the gap volume is measured by taking into account all the remaining allowable gap-spheres. Gap index calculated using the gap volume is a valuable method to evaluate complementarity between the interacting protein subunits. The gap index for different types of PPI complexes has been reviewed comprehensively by Jones and Thornton [23].

The gap index in a PPI complex is calculated as below:

$$\text{Gap index (Å)} = \frac{\text{gap volume between the molecules (Å}^3\text{)}}{\text{interface ASA (Å}^2\text{) (per complex)}} \qquad (2)$$

#### *1.4.3.1.10 Interface planarity*

Interface planarity is defined as the atomic root mean square deviation (RMSD) of all interface atoms from the least squares plane fitted through all interface atoms [40]. The PROTORP server shows an average interface planarity of ~3.1 Å for a dataset of 534 homodimers and 220 heterodimer complexes [155].

#### *1.4.3.1.11 Interface electrostatic potential*

The electrostatic interactions are the interactions between charged atoms or molecules. The negatively charged residues, aspartic acid (D) and glumatic acid (E) often form bonds with the positively charged residues, histidine (H), lysine (K) and arginine (R). It has been observed that electrostatics could drive the formation of interfaces, while specificity might be driven by interactions such as h-bonds, salt bridges and interactions between hydrophobic patches [190]. Significant population of charged and polar residues have also been observed on protein–protein interfaces [39, 191, 192].

#### *1.4.3.1.12 Interface hot spots*

A few interface residues are energetically more involved in the formation of interfaces than others called ‘hot spots’ [193-197]. Polar residues were observed to be conserved as hot spots in a dataset of 1629 protein-protein interfaces [198]. However, hot spots are structurally conserved and not conserved at sequence levels [197, 199]. Polar residues are known to occur at interfaces and are proportional to residue conservation, and size of the interface [182]. Furthermore, identification of hot spots using PPI features such as solvent accessible surface area, residue conservation and residue potential improves prediction accuracy for interface hot spots [199].

#### *1.4.3.1.13 Other properties of PPI interfaces*

Some PPI features besides those described above have also been used to identify and characterise PPIs. The differences in interface structural properties such as residues, hydrophobicity, hydrophilicity, hydrogen bonds, electrostatics, areas, and residue packing, steric strains, with selective pressure, has been documented for protease-inhibitors and antigen-antibodies [36]. Secondary structures at the biological interfaces with similar structural features between monomers was observed to contribute to the formation of a heterodimer complex [200]. van der Waals interactions along with other structural properties such as hydrogen bonds and hydrophobic interactions have been used to study PPI and differentiate among complexes [201]. Conformational changes upon protein complex formation have also been documented [202-204].

Analysis of structural features from known protein complexes helps in better understanding of features responsible for binding and extrapolating them to sequence level.

The various structural PPI features used in this thesis are analysed in Chapters 3 and 4.

### 1.4.3.2 Sequence-based features of PPI

Interface features based on non-structural data is also widely used for PPI analysis and prediction. Protein complexes are known to accommodate variation at sequence level, even with structurally similar interfaces [46]. Moreover, prediction of structural features from the sequence could improve sequence or evolutionary based prediction methods [205].

#### *1.4.3.2.1 Amino acid composition*

Amino acid composition is widely used as an important PPI feature to understand the chemical nature of the interface. Miller and colleagues showed that in small globular proteins the interface consists of 57% non-polar residues, 24% neutral polar residues and 19% charged residues [35]. However, the amino acid compositions in oligomeric proteins is made up of 65% non-polar residues, 22% neutral polar residues and 13% charged residues [206]. Therefore, the amino acid compositions differ among different complexes.

Aliphatic and aromatic residues as well as proline are the largest contributors to the interface as shown by Lijnzaad and Argos [51]. Non-polar residues are known to be more abundant in larger interfaces, while polar residues are abundant in smaller interfaces [207].

Lo Conte and colleagues showed that interfaces are rich in aromatic residues histidine, tyrosine, phenylalanine, and tryptophan than the average protein surface and somewhat richer in aliphatic residues leucine, isoleucine, valine, and methionine. However, they are depleted in charged residues except arginine [162]. The interfaces in hetero-complexes consisted of dominantly hydrophilic/polar residues (W, C, H, Q, N, Y, S, except T), as opposed to dominant hydrophobic/non-polar residues at the homodimer interfaces suggesting protein association by hydrophilic/polar interactions in non-identical complexes, as observed by Zhanhua and colleagues [43]. Amino acid composition and residue-contact preferences alone can predict interaction types with 63-100% accuracy as shown by Ofran and Rost [140].

#### *1.4.3.2.2 Amino acid residue propensity*

The interface amino acid residue propensity is the ratio of amino acids contributing to the interface as opposed to the ratio of amino acids contributing to the surface of a protein complex. This feature was first used by Jones and Thornton to study PPIs [23]. Their results showed that charged and polar residues, especially arginine and aspartic acid, show an increased affinity at the interface, and also non-polar residues such as methionine and proline show an increased affinity for the interface. Tryptophan and tyrosine residues are shown to have the highest propensity for the core of recognition sites, while serine and threonine have a negative propensity [150]. Bahadur and colleagues [41]  described interface features (residue propensity score) in specific and non-specific complexes using a dataset of 70 protein-protein complexes and 122 homodimers. Neuvirth and colleagues [208] counted tyrosine, methionine, cysteine and histidine as the most favoured residues at the interface, while threonine, proline, lysine, glutamic acid and alanine were least favoured at the interface. Gromiha and colleagues [172] showed dominance of aromatic and positively charged residues at the interface. Most studies show arginine as the major contributor to binding interfaces [182].

#### *1.4.3.2.3 Amino acid residue conservation*

Residue conservation at the interface is observed to be less than the core but more than the surface [153]. Chakrabarti and colleagues [41, 209] have also discriminated the interface into core and rim regions and have concluded that the core is relatively more conserved than the rim. Structurally conserved residues were also known to distinguish between binding interfaces and protein surfaces [195]. Their results showed that tryptophan residue conservation on a protein surface shows a highly potential binding site, while conservation of phenylalanine and methionine also implies a slightly potential binding site. Moreover, Guharoy and Chakrabarti [156] observed that conserved interface residues are not randomly distributed but distinctly clustered along the protein binding interface. Sequence and structure-based features have been used to analyse PPIs [210] for their utilization in training prediction models, however, no strict set of features is known as yet for accurate protein binding site predictions. The correlation between different features is observed to provide subtle differences [30]. Furthermore, 3D structures of protein-protein complexes gives an in-depth understanding of how two proteins interact physically [211]. Therefore, analysing existing protein-protein complexes and documenting their physicochemical features is important in gaining knowledge on the fundamentals of protein-protein association. Documenting these essential PPI features contributing to binding is hence necessary to mine a comprehensive set of features that are contributing to protein-protein recognition.

A combined formulation of structural and sequence-based features along with strong experimental evidence is therefore essential in understanding the molecular principles of PPIs and has applications in improving prediction accuracies of PPI prediction algorithms/models.

### 1.4.4 Protein interface databases

Protein interactions data has been deposited in numerous databases (Table 1.3) using interface datasets at protein and domain levels. PIBASE is a relational database of structurally defined interfaces obtained from protein domain pair interactions [212]. The database contains interfaces that are annotated based on geometric, physicochemical and topological properties responsible for the structural characterisation of protein complexes.

**Table 1.3: A list of currently available protein interface databases.**

| Database | Number of entries |
|---|---|
| 3did | 8829 domain-domain interactions; 291052 structures for domain-domain interactions |
| InterPare | 10,583 (Geometric distance), 10,431 (ASA), and 11,010 (Voronoi diagram) entries in PDB containing interfaces. |
| PDBsum | 111,180 entries including 2,196 superseded |
| PIBASE | 104,569 structures; 49,295 structures (PDB); 55,274 structures (PQS [216]); 598,638 domains; 212,071 domains (SCOP v1.73 [217]); 191,915 domains (CATH v3.1.0 [218]); 194,652 domains (chain); 755,998 interfaces; 269,821 interfaces (SCOP v1.73); 269,438 interfaces (CATH v3.1.0); 216,739 interfaces |
| PROTCOM | 1770 entries |
| SCOPPI | 105,547 domain-domain interactions; 22,874 domain-domain interactions at 90% non-redundancy level; 4,630 family-family interactions; 15,058 interface types |
| SCOWLP | 74,907 protein interfaces; 2,093,976 residue-residue interactions from 60,664 structural units |
| STRING | 5,214,234 proteins from 1133 organisms |

SCOPPI (Structural Classification of Protein-Protein Interfaces) database consists of classifications and annotations of protein domain interactions derived from PDB and SCOP domain definitions [213]. SCOPPI classifies domains based on their geometry and provides several interface characteristics such as number of interfaces, aminoacid types and positions, conservation, interface size, and permanent or transient nature of interactions. The database of 3D Interacting Domain (3did) archives domain-domain interactions obtained from high resolution protein structures [214, 215]. 3did provides information on structural similarity between different members of the same protein family.

PDBsum [115, 219] is a pictorial database providing summary information of 3D structures deposited at the PDB. The database offers structure diagrams, GO annotations, 1D sequence annotated by Pfam [220] and InterPro [221, 222] domain assignments, clefts in structures and schematic diagrams of PPI, including provisions for generating PDBsum information for user's own PDB file formats [115]. The SCOWLP (Structural Characterisation of Water, Ligands and Protein) web-server permits comprehensive structural analysis and comparisons of protein interfaces at atomic-level by text query of PDB codes and/or by navigating a SCOP-based tree. A visualization tool is involved to for users to interactively display protein interfaces and label interface residues and interface solvent by atomic physicochemical properties. SCOWLP is automatically updated with every SCOP release [223].

STRING [132] provides protein interaction interface information derived from high-throughput data, experimentally determined structures at PDB and literature. A large-scale protein domain interaction interface database, InterPare, is a public database and server for protein interface information obtained from 3D structures [124]. The database detects interfaces by three methods i.e. calculating geometric distance method for checking distance between atoms in different domains, detecting Accessible Surface Area (ASA) and calculating Voronoi diagram which uses mathematical definition for interaction interfaces. PROTCOM is a database of protein complexes enhanced with domain–domain structures [224]. Single chain structures are parsed into loosely connected domains to generate domain-domain structures. PROTCOM can be used as a template database to model 3D structures of unknown protein-protein complexes using homology modeling techniques or threading methods. Moreover, integrated set of tools for browsing, searching, visualizing and downloading a pool of protein complexes is also provided. PINT (Protein-protein Interactions Thermodynamic Database) [225] consists of >1500 data of thermodynamic parameters along with sequence, structural, experimental and literature information.

### 1.4.5 Protein interaction characterisation and prediction tools/servers

Protein interfaces are characterised by several webservers and/or tools based on their physicochemical properties. These webservers/tools developed over the past two decades for effective characterisation and prediction have been described in Table 1.4.

**Table 1.4: A list of webservers/tools developed for protein interface characterisation and prediction.**

| Webserver/Tool | Usage and URL |
|---|---|
| ConSurf | Identifies functionally important regions on the protein surface based on evolutionary conservation scores of protein residues<br>http://consurf.tau.ac.il/ |
| InterPreTS | Predicts protein interactions in query sequences using protein sequence similarity<br>http://www.russelllab.org/cgi-bin/tools/interprets.pl |
| InterProSurf | Predicts potential interacting amino acid residues on protein surfaces that are most likely to interact with other proteins.<br>http://curie.utmb.edu/ |
| LIGPLOT | Generates schematic 2-D representations of protein-ligand complexes with intermolecular interactions and their strengths, including hydrogen bonds, hydrophobic interactions and atom accessibilities<br>http://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/ |
| MAPPIS | Recognizes spatially conserved chemical interactions shared by a set of PPIs: http://bioinfo3d.cs.tau.ac.il/MAPPIS/ |
| meta-PPISP | A metaserver built on three individual web servers: cons-PPISP, PINUP, and Promate for predicting protein-protein interaction sites<br>http://pipe.scs.fsu.edu/meta-ppisp.html |
| MODTIE | Predicts binary protein interaction based on similarity of query sequence http://pibase.janelia.org/modtie/v1.11/index.html |
| PIC | Computes various interactions such as disulphide bonds, interactions between hydrophobic residues, ionic interactions, hydrogen bonds, aromatic–aromatic interactions, aromatic–sulphur interactions and cation–π interactions within a protein or between proteins in a complex: http://pic.mbu.iisc.ernet.in/job.html |
| PI2PE | Predicts protein interfaces http://pipe.scs.fsu.edu/ |

| Webserver/Tool | Usage and URL |
|---|---|
| PINUP | Prediction of protein binding site prediction<br>http://sparks.informatics.iupui.edu/PINUP/ |
| ProMate | Identifies the location of protein-protein binding sites<br>http://bioinfo.weizmann.ac.il/promate/ |
| ProtorP | Calculates physical and chemical parameters of the protein interaction sites such as size and shape, intermolecular bonding, residue and atom composition and secondary structure contributions<br>http://www.bioinformatics.sussex.ac.uk/protorp |
| Pred-PPI | Predicts PPIs from different organisms<br>http://cic.scu.edu.cn/bioinformatics/predict_ppi/ |
| ProFace | A server for the analysis of the physicochemical features of protein-protein interfaces<br>http://www.boseinst.ernet.in/resources/bioinfo/stag.html |
| ProteMot | Predicts protein binding sites based on the interaction templates automatically extracted from the compound crystals in the PDB<br>http://protemot.csie.ntu.edu.tw/step1.cgi<br>http://bioinfo.mc.ntu.edu.tw/protemot/step1.cgi |
| PRICE | Analyse protein-protein interfaces<br>http://www.boseinst.ernet.in/resources/bioinfo/stag.html |
| PPA-Pred | Predicts binding affinity of protein-protein complexes based on functional classification: http://www.iitm.ac.in/bioinfo/PPA_Pred/ |
| PPI Prediction Server | Discriminates PPI complexes into permanent, transient and crystal artefacts: http://ppi.zbh.uni-hamburg.de/ |
| SNAPPI | Calculates protein–protein interactions properties and is currently being employed to train a protein–protein interaction predictor and a functional residue predictor<br>http://www.compbio.dundee.ac.uk/SNAPPI/index.jsp |

| Webserver/Tool | Usage and URL |
|---|---|
| SHARP2 | Predicts potential protein–protein interaction sites on protein structures. http://www.bioinformatics.sussex.ac.uk/SHARP2 |
| SPPIDER | Predict residues at the putative protein interface(s) by considering single protein chain with resolved 3D structure; (2) analyse protein-protein complex with given 3D structural information and identify residues that are being in inter-chain contact.<br>http://sppider.cchmc.org |
| WHISCY | Predict protein-protein interfaces based on conservation and structural information<br>http://nmr.chem.uu.nl/Software/whiscy/index.html |

InterPreTS (Interaction Prediction through Tertiary Structure) developed by Aloy and Russell (2003) [226] uses BLAST to find homologues of known structure for all pairs for a given protein sequence set and therefore predicts protein interactions in query sequences using similarity between protein sequences and known complexes. MODTIE [227] predicts binary protein interactions and higher-order protein complexes from a set of protein sequences based on their similarity to template complexes, with the structures at MODBASE [228]. SHARP2 server [229], a flexible web-based bioinformatics tool, performs PPI prediction using patch analysis with parameters such as solvation potential, hydrophobicity, accessible surface area, interface residue propensity, planarity and protrusion.

The Proface server [179] developed by Saha and colleagues dissects PPI to derive physical and chemical features. Sppider [55] is a protein interface identification and recognition server based on solvent accessibility and structural information with high levels of accuracy (74%). The ProtorP server [155] predicts physical and chemical characteristics of structural interfaces (types of atoms at interfaces, structural elements, H-bonds, bridging water molecules, salt bridges and disulphide bonds, interface area, planarity, eccentricity, gap volume and gap index etc. in homo and hetero complexes), to assess the different interface properties of the query protein against the pre-calculated set of interface features in different complexes.

Some other webservers developed based on structural data of available protein complexes at the PDB for the characterisation and prediction of interaction sites and interfaces include cons-PPISP [53], meta-PPISP [230], PI2PE [231], PIC [232], LIGPLOT [233], PPI-Pred [234], PPI Prediction Server [235], MAPPIS [236], WHISCY [237], SNAPPI (Structures, iNterfaces and Alignments for Protein-Protein Interactions) [238], Promate [208], PINUP [239], ProteMot [240], ConSurf [241], InterProSurf [242], PepSite [59], PRICE [243], PPA-Pred [244], PrISE [58] and Wiki-pi [245].

**Table 1.5: List of webservers/tools developed for the prediction of hot-spots.**

| Webserver/Tool | Usage | URL |
|---|---|---|
| HotPoint | Analyse any protein–protein interface for binding sites characterization and rational design of small molecules for protein interactions. Identifies hot spots in protein interfaces by combining solvent accessibility and inter-residue potentials | http://prism.ccbb.ku.edu.tr/hotpoint |
| HotSpot Wizard | Identifies 'hot spots' for engineering of substrate specificity, activity or enantio-selectivity of enzymes and for annotation of protein structures | http://loschmidt.chemi.muni.cz/hotspotwizard/ |
| HSPred | predict hot spot residues based on support vector machine (SVM) method | http://bioinf.cs.ucl.ac.uk/hspred |
| KFC Server | Predicts binding "hot spots" within protein-protein interfaces by identifying structural features indicative of binding contacts | http://kfc.mitchell-lab.org/ |
| PredHS | Predicts PPI hot spots by using structural neighbourhood properties | http://www.predhs.org |

Databases and webservers/tools available for predicting protein-protein interface hot spots including HotPoint [199, 246], HotSpot Wizard [247], HSPred [248], KFC server [249] and PredHS [250] as given in Table 1.5.

Other databases for hot spots prediction include ASEdb [251], BID [138], HotSprint [252], APIS [253], PCRPI [45], and DBAC [254]. PPI hot spots and *in silico* techniques available for hot spot residues is reviewed by Morrow and Zhang [47].

On the other hand, community wide experiment, CAPRI (Critical Assessment of Predicted Interactions) [255], performs continuous assessments of prediction models several times in a year. The prediction of interaction regions and residue contacts while accommodating large conformational changes with accuracy is an non-trivial task [256]. The community experiment recently found that electrostatics and solvation terms marginally distinguish the designs of proteins from natural complexes, largely due to non-polar characteristics of interactions and also that the binding surfaces were structurally less embedded in designed monomers, suggesting the importance of conformational rigidity at the designed surface [257].

## 1.5 INTRODUCTION TO INTEGRINS

Integrins are transmembrane glycoproteins that mediate interactions between the components of extracellular and intercellular milieu. Integrins are found in multicellular organisms, from sponges to mammals [258, 259]. These receptors span the plasma membrane promoting anchorage across the extracellular matrix and the cell components. Bidirectional cell signalling of integrins transduces information between the components of extracellular and intracellular milieu [260]. The extracellular matrix binds to integrin glycoproteins initiating structural changes, subsequently triggering signal transduction [261, 262]. These signals (arising from receptor-binding) relate to cell migration, attachment, differentiation, proliferation, polarity and survival/apoptosis [261, 263].

Integrins are also known to regulate biological processes related to cell morphology, proliferation, survival, migration and invasion mostly by involving in crucial cell signalling pathways [264]. These receptors are involved in initiation and/or progression of many malicious diseases including tumour metastasis, immune dysfunction, neoplasia, inflammation, trauma and infections as reviewed elsewhere [260, 265-269]. These receptors have been the target of therapeutic drugs to combat inflammation, thrombosis, fibrosis and tumourigenesis caused by many viruses and bacteria [266, 270-273].

The integrin structure consists of two distinct α and β chains forming heterodimers with an obligatory function. These α and β subunits assemble into a "head" segment built on top of two V-shaped "legs" [274]. In mammals, 18α chains and 8β chains have been characterised for integrins which noncovalently associate to form 24 different receptors [266, 275]. These 24 receptors binding to specific ligands have been characterised to have a unique function as shown in Figure 1.14. Ligands binding to the integrins include fibronectin, vitronectin, collagen, laminin and cytotactin. Individual integrins specifically bind to protein ligands, while the R-G-D (arginine-glycine-aspartate) tri-peptide sequence is a commonly known integrin-binding motif [276] as shown in Figure 1.14.

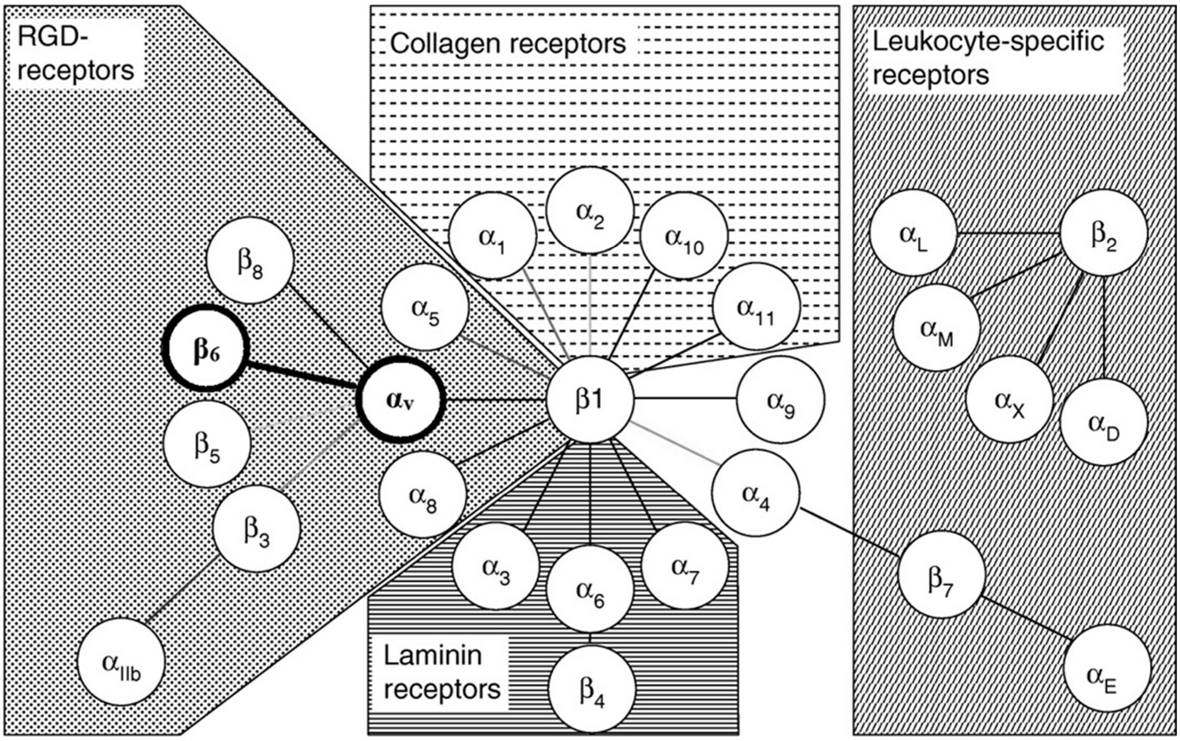


**Figure 1.14: Integrin receptor classes and related integrin-targeted compounds.** The various classes of integrin heterodimer receptors are shown. The integrin consists of specific α and β binding subunits for functionality. The β6 subunit binds exclusively to αv subunit. Adapted from Binder and Trepel, 2009 [8].

Integrins are made of a comparatively large extracellular domain, a transmembrane domain and a short cytoplasmic tail [258, 266]. The α and β subunits of an integrin structure have an amino acid length over 1000 and 750 residues, respectively. The integrin β subunit is recognized to play a key role in regulatory function [277]. The β subunit undergoes conformational changes on ligand recognition resulting in variable epitope expression levels, thereby regulating integrin activity [278]. Elucidation of three-dimensional (3D) structures of integrins [262, 279] in the past decade, has paved way for researchers to perform intensive structural analyses to relate to the functional significance of these large glycoproteins. However, the structural basis of integrin activation and regulation is yet to be known [261].

### 1.5.1 Integrin αvβ6 heterodimer

Integrin αvβ6 is an epithelial heterodimeric transmembrane protein comprising a β6 subunit which binds exclusively to an αv subunit [280, 281] as shown in Figure 1.14. The αvβ6 integrin expression is primarily restricted to epithelial cells, where it is expressed at low-levels in normal adult cells while elevated in fetal tissue during embryogenesis, morphogenesis, and in injured tissue during wound healing, inflammation and tumourigenesis [280, 282-284]. Internalisation of integrin αvβ6 via clathrin-mediated endocytosis promotes cancer cell invasion. The αvβ6 integrin mediates cell adhesion, proliferation, migration and invasion. Integrin αvβ6 binds to ligands including fibronectin, cytotactin/tenascin, vitronectin, and Transforming Growth Factor-β1 (TGF-β1). Activation of TGFβ1 by the αvβ6 integrin occurs as a result of ligand-binding, through interactions with the R-G-D motif present in the latency associated peptides 1 and 3 (LAP1 and LAP3) [281, 285, 286]. These heterodimers are also known to be involved in the activation of phospho-ERK2, MAP kinase pathway and TGF-β1 pathway at the cell surface [280]. Furthermore, the interaction of integrin αvβ6 to HAX-1, a HS1 associated protein [287] involved in endocytosis, plays a major role in cancer progression [288], by controlling the net signalling output of the cell at various stages, thus associating this integrin in various stages of cancer.

The unique 11-aminoacid C-terminal cytoplasmic tail of the integrin β6 subunit mediates cell proliferation, matrix metalloproteinase production (MMP2 or MMP9), invasion and survival [289]. Mutations in the cytoplasmic domain of the β subunit is identified to affect integrin activity [290]. Aminoacid extensions of the integrin β6 subunit have specific interactions with the cytoplasmic components, distinguishing it from its closely related homologue, integrin β3 [291].

The αvβ6 integrin is also known to have a role in providing immune tolerance [292]. Neutralizing the αvβ6 is assumed to disrupt malignant transformation of cells [293]. The over-expression of αvβ6 integrin in a number of epithelial cells is hypothesized to promote malignancy through alterations in cell proteolytic activities with implications in cancer metastasis [294]. High levels of integrin expression have been documented in various epithelial carcinomas and cancer cell lines including skin, oral, lung, breast, pancreas, liver, gastric, ovary, basal cell, endometrium, cervical squamous, duodenal and colorectal adenocarcinomas [282, 283, 293-307]. Interestingly, αvβ6 regulates development, neoplasia, and tissue repair, suggesting a role in epithelial remodelling [283].

### 1.5.2 The urokinase plasminogen activator receptor protein (uPAR)

The uPAR protein is a versatile signalling orchestrator mediating interactions with other transmembrane receptors, including integrins. It is a glycosylphosphatidylinositol (GPI) anchored extracellular membrane protein. uPAR acts as a specific receptor for the urokinase plasminogen activator (uPA). In 1985, the uPAR was first identified in monocytes and monocyte-like U397 cells [308]. Subsequent studies on uPAR confirmed its involvement in many cell signalling pathways. uPAR is known to have implications in cell migration, adhesion, proliferation and tissue remodelling [309, 310]. uPAR is found on the surfaces of neoplastic and inflammatory cells including circulating blood monocytes and neutrophils [311, 312] in a normal cell. However, the expression of uPAR is high during cancer progression. Thus, in normal biology, uPAR plays a key role in enhancing extracellular proteolysis through confining plasminogen activation though uPA proximal to cell surface [313, 314]. uPAR is known to play a role in various types of cancer including breast, ovary, colon, lung and other carcinomas [315-317].

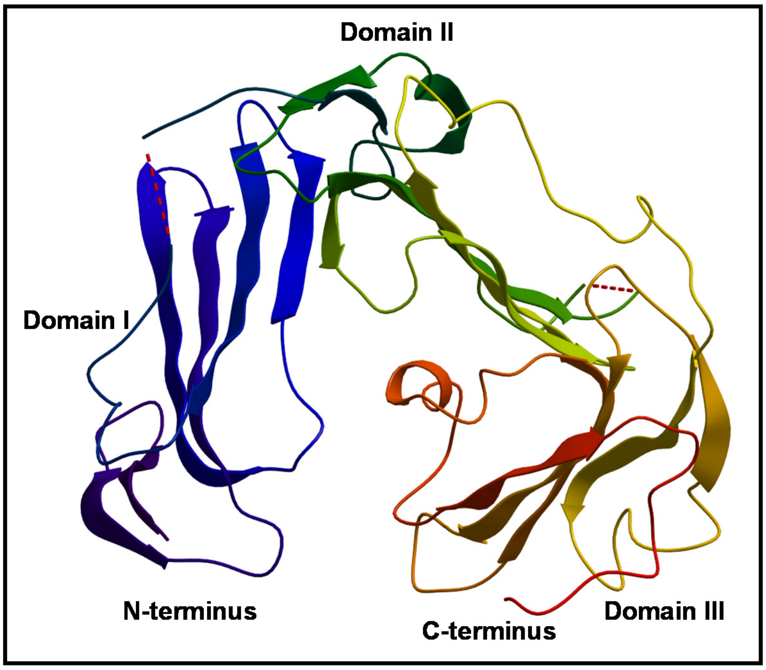


**Figure 1.15: The human uPAR structure is shown.** The structure of human uPAR is shown in ribbon representation with three domains Domain I, Domain II and Domain III.

The uPAR protein forms a glove-like structure (Figure 1.15) providing a central pocket for the binding of uPA [318]. It is made up of a single chained polypeptide with a sequence length of 283 amino acids. The three extracellular domains DI, DII and DIII are Ly6/uPAR/α-neurotoxin-like (LU) domains consisting of approximately 90 amino acid residues each [319]. The uPAR has a three polypeptide loop structure with each domain having four to five disulphide bonds [320]. Each of these domains is connected to the other by a linker region (i.e. L1 and L2) and has six anti-parallel β- strands, while DIII alone has five β-strands, ending with two short helical stretches. The domains adopt a three 'finger' fold to form a concave cavity in the centre of the receptor (Figure 1.15), with a high affinity for uPA.

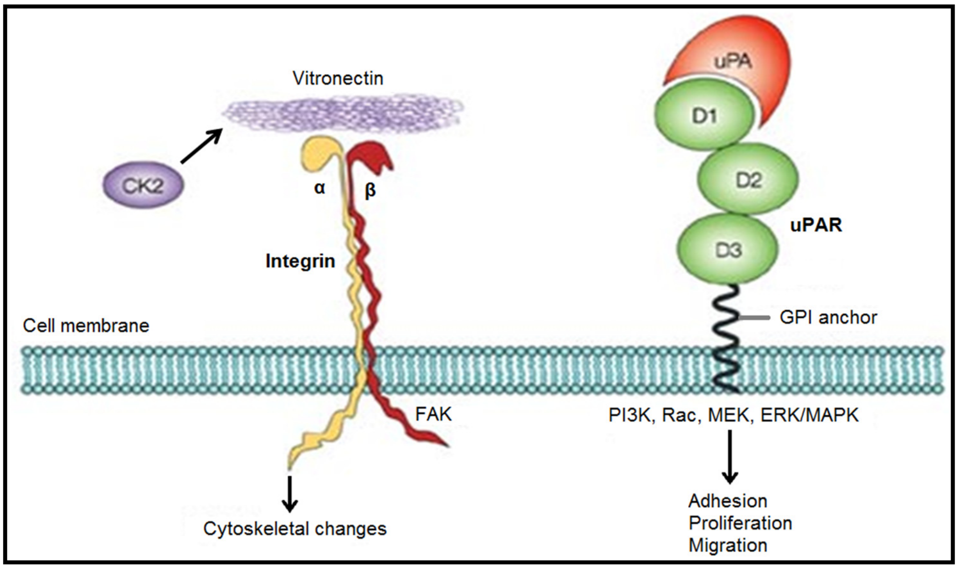


**Figure 1.16: Schematic showing the role of uPAR•integrin interaction.** Adapted from Blasi and Carmeliet, 2002 [9].

### 1.5.3 uPAR•integrin interaction

uPAR is believed to play a role in downstream cellular signalling pathways through lateral interactions with transmembrane proteins such as integrins as they lack intrinsic intracellular domains [321]. Figure 1.16 illustrates the role of uPAR•integrin interaction in activating various pathways leading to biological functions such as adhesion, proliferation and migration, within the cell. Binding of integrin αvβ6 to urokinase plasminogen activating receptor (uPAR) promotes the plasminogen activator system, thereby playing a key role in cancer progression [322]. An in-depth understanding of the PPIs involving integrin αvβ6 and uPAR is essential to understand the uPAR•integrin interaction for specific activity. Hence, I have carried out a PPI study on integrin αvβ6•uPAR interactions using modeling data and docking simulations as detailed in Chapter 5.

## 1.6 OBJECTIVES

PPIs form the central basis for complex biological networks and molecular functions in a living cell. These interactions play a key role in the fields of systems biology, functional genomics and drug design. Therefore, there is a need to comprehensively study these interactions to gain insights into their molecular principles for applications in PPI prediction and identifying targets for drug design. Experimental determination of protein structures and interactions using high-throughput techniques has made it possible to obtain X-ray crystal structures of protein and also valuable information pertaining to protein interactions at cellular level. Bioinformatics analyses can help obtain into the binding principles of these interactions using the experimentally determined protein structures. A membrane heterodimer complex involved in cancer progression has been comprehensively studied to evaluate what properties characterize PPIs. Based on these properties, complexes were grouped at residue and functional levels to obtain distinguishing features for applications in prediction models. Specific aims are listed below, with four publications presented in this thesis:

1. Review the key resources available for the study of PPIs, experimental procedures for PPI determination, the computational methods in determining PPIs, the key databases archiving this information, current trends involved in studying PPIs followed by key datasets created/collected by various groups and the deterministic PPI features known to govern PPIs (Publication 1).
2. Structural analysis of all non-redundant known heterodimeric protein complexes at the PDB and classifying them based on their residue-level relative interface-surface polarities to understand predominant interactions at the interface and identify discriminatory PPI features between these classes (Publication 2)
3. Group protein complexes based on literature-driven molecular functions to identify structural features possibly characterising functional interfaces (Publication 3).
4. Study integrin αvβ6 membrane protein using model data and docking simulations for the characterisation of integrin αv-β6 subunit interface to gain insights into the structural basis of integrin αvβ6•uPAR interactions (Publication 4).

## *1.7 Publication 1*

# Protein-Protein Interactions and Prediction: A Comprehensive Overview

Gopichandran Sowmya[1] and Shoba Ranganathan[1,]*

[1]*Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia*

**Abstract:** Molecular function in cellular processes is governed by protein-protein interactions (PPIs) within biological networks. Selective yet specific association of these protein partners contributes to diverse functionality such as catalysis, regulation, assembly, immunity, and inhibition in a cell. Therefore, understanding the principles of protein-protein association has been of immense interest for several decades. We provide an overview of the experimental methods used to determine PPIs and the key databases archiving this information. Structural and functional information of existing protein complexes confers knowledge on the principles of PPI, based on which a classification scheme for PPIs is then introduced. Obtaining high-quality non-redundant datasets of protein complexes for interaction characterisation is an essential step towards deciphering their underlying binding principles. Analysis of physicochemical features and their documentation has enhanced our understanding of the molecular basis of protein-protein association. We describe the diverse datasets created/collected by various groups and their key findings inferring distinguishing features. The currently available interface databases and prediction servers have also been compiled.

**Keywords:** Binding sites, interface features, prediction, protein complexes, protein-protein interactions.

## INTRODUCTION

Proteins are social molecules essential for several biological processes in a cellular system. Proteins bind to different molecules such as organic and inorganic compounds, metals, sugars, fatty acids, nucleotides and other proteins. Protein-protein interactions (PPIs) form the central basis of complex cellular networks, thereby playing a pivotal role in the fields of systems biology, functional genomics and drug design [1-4]. PPIs occur with varying affinities, yet with a high degree of specificity [5]. Specific interactions forming stable interfaces between two or more protein molecules lead to diverse functions such as catalysis, regulation, signal transduction, immunity and inhibition [6, 7]. Hence, a study of the interactions between specific proteins is the key to understanding cellular machinery.

Progress in large-scale high-throughput experimental procedures has led to the determination of possible PPIs within an entire genome. However, these biological experiments are laborious, time-consuming and expensive, besides their major drawback of poor accuracy in data generation and the inclusion of many false positive proteins that are not necessarily associated *in vivo* [8-11]. Moreover, a comparative assessment of large-scale datasets of PPI has also shown that *in silico* predictions of many interactions provided levels of accuracy close to those determined experimentally [9]. Therefore, the flaws involved in the experimental determination of PPI complexes pose a need for progress in computational methods that can precisely reflect biological reality from abstract structural data [11, 12].

At the other end of the spectrum, protein-protein docking procedures are being used for the prediction of native conformations of multimeric proteins when the constituent protein structures are known or by using high quality three-dimensional (3D) structural models, known as "targeted interaction pairs" [13, 14]. Docking methods can predict near-native conformation of the complex based on comparative modelling of known protein structures using structural features such as shape complementarity, steric, geometric and energetic considerations [15-19]. Nevertheless, docking studies lack high levels of accuracy owing to the inadequate information on the forces that bind proteins together [20]. Structural analyses of known protein complexes provide features that complement docking studies for accurate predictions of 3D structure of the PPI complex [21-24].

Protein-protein complexes have been exhaustively analysed for common features as an important step towards deciphering the binding principles and functionality of proteins [4, 25-27]. The ability of a protein to interact with its partner depends on several physicochemical features that are additive in nature. A stable interface is often formed between the interacting partners. Figure **1** represents a generic PPI complex, composed of a heterodimer, with interacting residues forming a stable interface. Protein-protein interfaces have several distinct features that distinguish them from the rest of the protein surface. Structural data of X-ray 3D protein complexes available at the Protein DataBank (PDB) [28] is commonly used to study protein-protein interfaces [29-31]. The interacting residues are characterized based on physical

*Address correspondence to this author at the Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia;
Tel: +61298506262; Fax: +61298508313;
E-mail: shoba.ranganathan@mq.edu.au

1875-5305/14 $58.00+.00

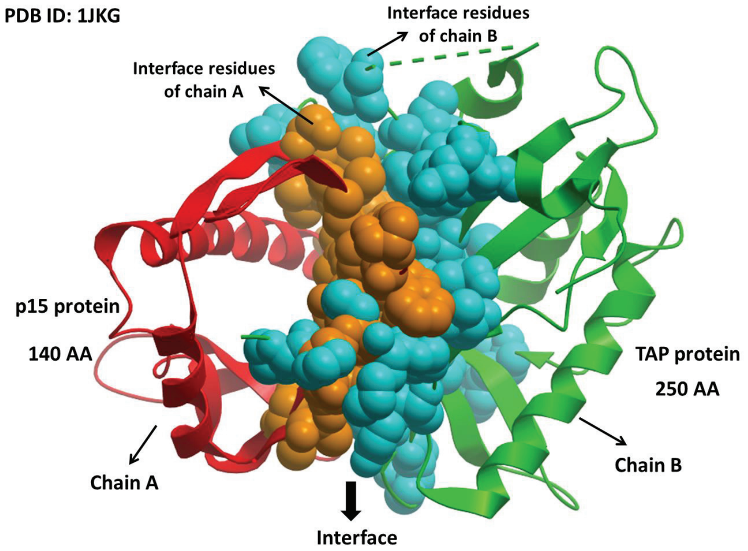


**Figure 1. A pictorial representation of a protein-protein interaction complex**. TAP-p15 heterodimer (PDB ID: 1JKG) is shown here. The interacting residues at the protein interface are shown in space-filling or CPK representation with interface residues of chain A and chain B coloured amber and cyan, respectively. The protein chains are shown in cartoon representation with chain A and chain B coloured red and green, respectively.

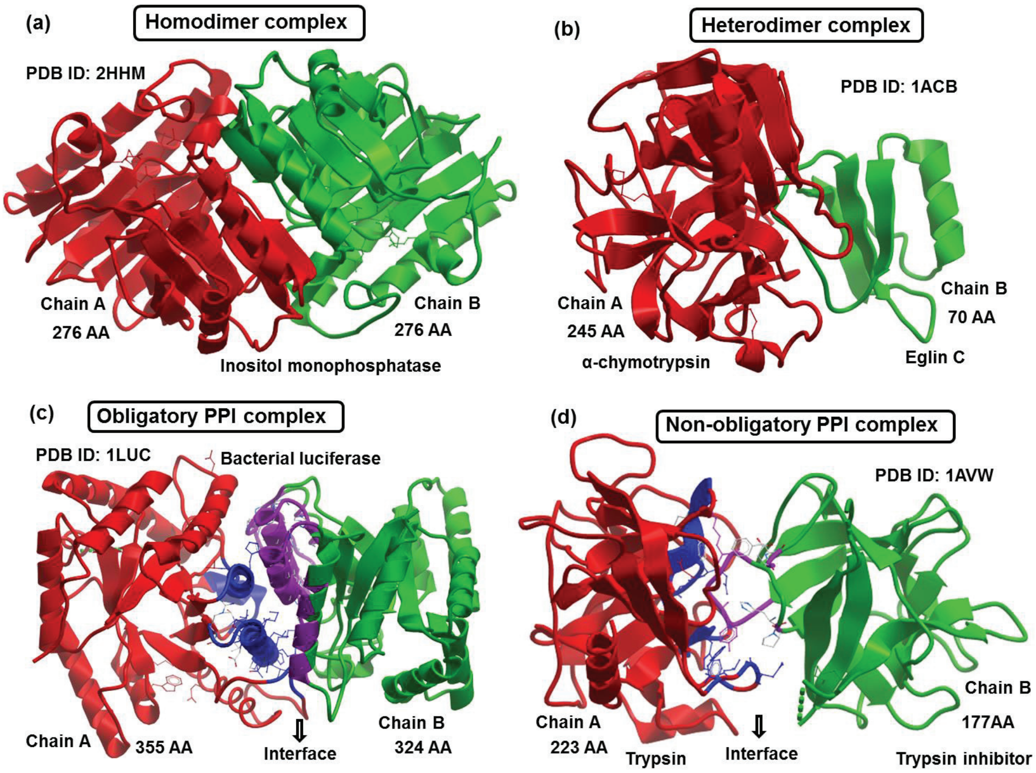


**Figure 2: Classification of PPIs shown with structures of example cases**. Chain A and chain B of each of these complexes are shown in red and green respectively. (**a**) Homodimer with the two-identical chains (chain A and B) forming the inositol monophosphatase complex performing catalyses in phosphatidylinositol signalling pathway; (**b**) Heterodimer with two non-identical chains (Alpha-chymotrypsin and Eglin C) performing an enzyme-inhibitory function; (**c**) Obligatory complex of bacterial luciferase with chain A and chain B forming a stable interface (coloured in blue and magenta, respectively) to catalyse the oxidation of long-chain aldehydes; (**d**) Non-obligatory complex of pancreatic trypsin/trypsin inhibitor interacting transiently at the interface coloured as in (**c**).

and chemical features based on their strengths of interactions in different types of complexes.

Proteins are also known to interact with multiple partners forming many interfaces [32, 33]. The formation of several interfaces among protein multimers is yet another facet of PPI that poses difficulty in prediction efforts, possibly because the interaction between the multiple protein partners occurs at different levels and these associations are often weak or transient in nature [34]. Clearly, the paucity in the availability of 3D structures of such transient complexes in the PDB hampers the inclusion of such interfaces for the analyses of binding sites and partners. Dimer complexes are known to be the strongest and most extensive interactions in nature, as their individual isolated oligomer subunits rarely exist as functional monomers [26]. Therefore, structural analyses of available protein complexes aids in better understanding the molecular basis of protein-protein association.

Here, we present an overview of the experimental methods used to determine PPIs, the key databases archiving these experimental data, as well as computationally predicted PPI information. We then review the current trends in interaction analyses and prediction, describing the classification of PPIs and differences in their interface features. The variations in PPI datasets created/collected by different groups, based on differing PPI features, are discussed. These datasets have led to the creation of interface databases that provide the data for currently available interaction characterisation and prediction tools/servers.

## EXPERIMENTAL DETERMINATION OF PPIs

Determination of PPI is based on a comprehensive characterisation of qualitative and quantitative aspects of PPIs [10]. However, experimental approaches are often incomplete, providing qualitative information to catalogue PPI without taking heed to the quantitative and dynamic features involved in such interactions. Moreover, experimental determination of PPIs for proteomes of entire species provided a smaller number of PPIs than the expected number or the number of interactions is often underestimated [35-37]. Computational approaches complement experimental data for an effective determination of PPI by predicting and prioritizing the data for experimental study, thereby reducing their costs and time consumed. Methods for measuring protein interactions in living cells have been previously reviewed by Piehler [38]. We provide an outline of the two broad categories of PPI determination approaches such as Fragment Complementation (FC) assays and affinity purification methods. FC is based on the functional reconstitution of proteins by fusion of the interacting proteins, thereby determining binary interactions between protein pairs and has been widely used in yeast organism. On the other hand, affinity purification methods combined with MS (Mass Spectrometry) perform structural determination of all the components in protein complexes, and have been used in several large-scale studies to investigate PPIs in model organisms and in human.

### Fragment Complementation Assay

The yeast two-hybrid (Y2H) is a widely used PPI determination technique, initially developed by Field and Song in 1989, taking advantage of the properties of GAL4 protein of yeast *S. cerevisiae* [39]. The GAL4 protein is a transcriptional factor, which activates the expression of a reporter gene when its DNA-binding domain (DBD) and its transcription activation domain (AD) are linked. However, GAL4 protein loses its capability of activation when the two domains (DBD and AD) are separated. In this technique, the two proteins of interest are fused with either domain (DBD and AD) of the transcription factor. Interaction between the proteins reconstitutes the functional form of the domains, thereby activating the expression of reporter gene. Capitalizing on the activation property of GAL4 protein, the Y2H technique effectively determines whether two proteins truly interact with each other. Various other FC techniques have been developed over the past few decades for the detection of PPI based on the co-expression of two-hybrid fusion proteins [40-47].

### Affinity Purification Methods

The principle behind affinity purification methods is that the interactions of protein partners involving affinity-tagged proteins formed *in vivo* are preserved during biochemical purification steps. In this technique the proteins of interest are purified from the cell and their interactions are identified *in vitro* under physiological conditions [48]. GST-pulldown and co-immunoprecipitation (co-IP) are the two widely used affinity purification techniques, supplemented with refined high-throughput methods that use mass spectrometry for protein identification. GST-pulldown technique uses glutathione S-transferase (GST) as a tag (often radio-labelled) for studying *in vitro* protein interactions. A number of studies also reported the interactions of proteins using Co-IP methods [49-51]. Co-IP uses Protein A (isolated from Staphylococcus aureus) as a tag for identifying interactions between proteins; however, these interactions are limited based on the availability of specific antibodies. The various experimental techniques available for deciphering protein-protein interactions, their advantages and disadvantages, and approaches to validate the diverse data produced by high-throughput techniques has been reviewed by Shoemaker and Panchenko [52].

### Databases for PPI

Protein-protein interaction data, determined as a consequence of experimental and bioinformatics approaches, is made available through large scale genome- and proteome-wide analyses. Several key databases archive the experimentally verified as well as computationally predicted interactions data. While several databases provide information on the physical interactions few others also provide information on the associations based on functional gene links and interactions at domain level [53]. Inclusion of protein associations at different levels as in multiple partner complexes, cellular components, metabolic pathways, co-expression, gene regulation or molecular co-evolution represents a confounding data landscape. Therefore, incorporating new data, organizing, curating, annotating and thereby linking the data objects to several biological characterizations are of utmost importance in PPI databases. A list of databases with their characteristics such as source organism, detection type, structural availability, interaction type and the number of interactions has been reviewed by Tuncbag and colleagues [53]. Some of the available PPI databases with updated information (as of February, 2013) include DIP (Database of Interacting Proteins) [54] and related databases [55, 56], BIND (Biomolecular Interaction Network Database) [57-59], MINT (Molecular INTeraction Database) [60], IntAct [61], BioGRID (Biological General Repository for Interaction Datasets) [62], HPRD (Human Protein Reference Database) [63] and STRING [64], as listed in Table **1**. Based on the specific tasks at hand, a combination of these databases has been widely used by researchers to obtain information or to validate their predicted interactions.

## CURRENT TRENDS IN COMPUTATIONAL PPI ANALYSIS AND PREDICTION

### Classification of PPI Complexes

Protein-protein interaction complexes can be grouped into several classes based on their composition, affinity and the lifetime of their association [65]. Non-identical protein subunits (or chains) interact to form hetero-oligomers while identical protein subunits (or chains) interact to form homo-oligomers. Examples of protein homodimer (with two identical chains) and heterodimer (with two non-identical chains) complexes are shown in (Figs. **2a**, **2b**). The folding and binding mechanism in the formation of homo-oligomers based on intra-molecular/inter-molecular contacts is intriguing, since homo-dimeric interfaces are formed through three folding mechanisms such as 2S (2S without any intermediate), 3SMI (3-state with monomer intermediate) and 3SDI (3-state with dimer intermediate) [66]. Therefore, it is often necessary to understand the homo-dimeric association and folding mechanism of known structural complexes to predict the folding mechanism given their structural complexes [67]. Compared to heterodimers, homodimers are known to have larger interfaces and more H-bonds, with predominant hydrophobic residues at their interfaces [68, 69] however heterodimers have a higher density of H-bonds per residue with predominantly charged and hydrophilic residues at their interfaces. The binding modes in homomeric complexes also play a role in distinguishing these from the heteromeric complexes, based on electrostatics [70]. Thus, the structural analysis of protein dimers (homo- and heterodimers) using known protein 3D complexes reveals significant differences in the interface features.

PPIs are also classified based on their affinity and lifetime into obligate or non-obligate and permanent or transient complexes, respectively [65]. The constituent protein subunits (or monomers) in obligate PPI do not exist as stable structures *in vivo*, while the constituent proteins in non-obligate PPI complexes can exist independently. Examples of obligatory and non-obligatory PPI complexes are shown in (Figs. **2c**, **2d**). Discrimination of PPI into permanent complexes existing in stable and irreversible complexed forms and transient complexes readily associating and dissociating to perform a functional activity *in vivo* is based on their lifetime of association. Permanent and transient complexes can be distinguished based on structural features such as interface geometrical and physicochemical properties and sequence properties such as aminoacid substitutions [27, 65, 71, 72]. Typically, obligate interactions are permanent/ stable, whereas non-obligate interactions are predominantly transient, although a few are permanent in nature [6, 65]. The analysis of six types of PPIs, such as interactions within the same structural domain and between different domains, per-

**Table 1. Current databases for PPI.**

| Database | Description | Number of entries | Reference |
|---|---|---|---|
| DIP (Database of Interacting Proteins) | PPI data determined by experimental techniques (yeast two-hybrid, protein microarrays and TAP/MS) | 75,019 interactions from 25,388 proteins and 541 organisms | [54] and related databases [55, 56] |
| BIND (Biomolecular Interaction Network Database) | Provides hand-curated molecular interactions data extracted from high-through experiments and literature | Over 206,859 unique interactions | [57-59] |
| MINT (Molecular INTeraction Database) | Relational database archiving PPI reported from peer-reviewed articles | 241,458 interactions from 35,662 proteins spanning over 30 organisms | [60] |
| IntAct | Molecular interaction database providing experimentally determined interactions across several species from literature curation and from user submissions | 305,970 binary interactions compiled primarily from 6177 publications and 17,905 experiments | [61] |
| BioGRID (Biological General Repository for Interaction Datasets) | Provides comprehensive curation of protein-protein and genetic interactions | 444,517 physical and genetic non-redundant interactions from 47,972 unique proteins | [62] |
| HPRD (Human Protein Reference Database) | Contains manually-curated protein interaction and pathways information of human proteins | 41,327 PPIs with 30,047 proteins. | [63] |
| STRING | Known and predicted protein interactions including direct (physical) and indirect (functional) associations obtained from genomic context, high-throughput experiments, (conserved) co-expression and previous knowledge | >5.2 million interactions from 5,214,234 proteins and 1133 organisms | [64] |

manent and transient interfaces, homo- and hetero-oligomers (oligomer is a polymer consisting of monomeric units) showed that aminoacid composition alone can be used to distinguish the different classes of complexes [73]. Therefore, it is essential to discriminate the different types of protein complexes for PPI analysis and prediction studies, to gain knowledge on their nature of binding and functionality [65, 74, 75].

### Datasets for PPI

Protein-protein interactions are often studied using a non-redundant dataset of structural complexes obtained from protein structural repository databases such as PDB [28]. The X-ray crystal complexes deposited each year at the PDB are growing exponentially. Obtaining a non-redundant yet reliable dataset of protein structural complexes from the PDB, for PPI analysis and to train predictors, often poses difficulty owing to the unorganized structure in data repository, despite significant efforts. The data deposited at the PDB have few true or real complexes as compared to crystal artefacts in spite of much progress in determining the 3D structure of proteins [74]. The paucity of biological data at the PDB poses difficulty in maintaining a default procedure to mine for a reliable dataset or in using a standard dataset for analysis and to train predictors. Therefore, creating an updated yet non-redundant dataset representing protein multimers from PDB is a non-trivial task and an important step in PPI studies [7].

Structural datasets created by different groups for PPI studies consists of heterogeneous data as listed in Table **2**. Protein-protein docking benchmark datasets [76] have also been extensively utilized to understand the roles of protein-protein interfacial residues in binding [77]. Several authors have also grouped the dataset based on the types of PPI to further study the bias in interface properties in different groups of complexes [65, 78, 79].

### Discerning Features of PPI

The protein interface formation between the two interacting subunits is governed by both physical and chemical features as described in a number of studies [25, 26, 29, 80-83]. Physicochemical features that govern the subunit interactions include shape complementarity, planarity, sphericity, interface size, interface area, gap volume, gap index, residue side-chain packing, residue propensities (frequency of amino acid residue type), electrostatics, hydrogen bonds, salt bridges and disulphide linkages. These features have been studied extensively using structural datasets of protein-protein complexes, to better understand the features determining their driving force for binding and thereby train predictors.

Interface features based on non-structural data is also widely used for PPI analysis and prediction. Chemical nature of the residues (amino acid composition) at the interface determines the hydrophobic effect of protein binding. Hydrophobic residues were predominantly observed at the interface [80]. Interface with significantly hydrophilic as well as hydrophobic characteristics similar to surface with few charged groups were also observed in a dataset of 75 complexes [29]. Ofran and Rost (2003) [73] have shown that amino acid composition and residue-contact preferences alone can predict interaction types with 63-100% accuracy. The interfaces in hetero-complexes consisted of dominantly hydrophilic residues (W, C, H, Q, N, Y, S, except T), as opposed to dominant hydrophobic residues at the homodimer interfaces suggesting protein association by hydrophilic interactions in non-identical complexes, as shown by Zhanhua and colleagues [69]. Residue conservation at the interface is observed to be less than the core but more than the surface [84]. Chakrabarti and colleagues [85, 86] have also discriminated the interface into core and rim regions and have concluded that the core is relatively more conserved than the rim. Prediction of structural features from the sequence could improve sequence or evolutionary based prediction methods [87].

Analysis of structural features from known protein complexes helps in better understanding of the features responsible for binding and extrapolating them to sequence level. Miller and colleagues (1987) [81] have shown that accessible and buried surface area is related to molecular weight. The differences in interface structural properties (residues, hydrophobicity, hydrophilicity, hydrogen bonds, electrostatics, areas, and residue packing, steric strains) with selective pressure has been documented among protease-inhibitors and antigen-antibody [25]. Number of residues forming the interface is proportional to the interface area [30, 85]. Stronger protein subunit binding was commonly associated with larger interface areas [26]. The stability of a protein-protein association also depends on the number of hydrogen bonds at the interface. Therefore, documentation of inter-molecular hydrogen bonds among different protein complexes helps determine stability and evaluate predictions [26, 27, 29, 85, 88]. Structural properties such as hydrogen bonds, hydrophobic and Van der Waals interactions have been used to study PPI and differentiate among complexes [89]. Interface patch is yet another feature widely studied to determine protein-protein recognition mechanism. Analysing the distribution of surface patches at the interface using structural features showed trends for distinguishing interfaces from other surface patches [68]. Dominant peptide segments were documented at the interfaces of homodimer and crystal contacts [90]. Conserved residues were observed to be distinctly clustered and not randomly distributed on the interface [91]. It should also be noted that protein complexes are known to accommodate variation at sequence level, even with structurally similar interfaces [5]. Documenting the essential structural features that directly or indirectly contribute to binding is hence necessary to mine a comprehensive set of features that are causal to protein-protein recognition [92].

Several sequence and structure-based features have been used to analyse and comprehend PPIs [93], and for their incorporation in training predictors, however, no strict set of features is known as yet for accurate binding site predictions. Also, the correlation between different features may provide subtle differences [74]. Moreover, a thorough understanding of how proteins physically interact with each other can be gained from 3D structural information [94]. Therefore, documenting the roles of various features that contribute to binding in different complexes is often essential in under-

**Table 2. PPI datasets and their key findings.**

| Group | Year | Dataset | Key findings | Reference |
|---|---|---|---|---|
| Chothia and Janin | 1975 | 3 (insulin dimer, trypsin-PTI, α/β oxy-haemoglobin) | Dominant hydrophobic interfaces | [80] |
| Chothia *et al.* | 1976 | 2 (horse methemoglobin, human hemoglobin) | Dimer-dimer interfaces are close packed and hydrophobicity stabilises the structure | [137] |
| Miller *et al.* | 1987 | 11 dimer, 9 tetramers, 2 hexamer, 1 octomer | Same molecular weight bury similar amounts of surface, while the proportions buried within and between subunits vary unlike monomeric proteins | [81] |
| Janin and Chothia | 1990 | 15 protease inhibitors, 4 antigen-antibody complexes | Interface structural properties and their implications for kinetics and thermodynamics of association | [25] |
| Jones and Thornton | 1995 | 32 complexes | Higher interface hydrophobic residues than surface but less than interior | [26] |
| Jones and Thornton | 1996 | 59 complexes | Explored factors influencing interface formation are among different complexes | [27] |
| Xu *et al.* | 1997 | 319 protein-protein interfaces | Studied hydrogen bonds and salt bridges for specificity of protein-protein associations | [88] |
| Tsai *et al.* | 1997 | 362 non-redundant protein-protein interfaces, 57 symmetry-related oligomeric interfaces | Interface hydrophobic residues involved (although not as much as in protein folding) in binding | [82] |
| Dasgupta *et al.* | 1997 | 58 oligomeric proteins, 223 protein crystal structures | Hydrophobic interactions at oligomeric interfaces favour aromatic amino acids while crystal contacts avoid inclusion of hydrophobic interactions. | [138] |
| Linzaad and Argos | 1997 | 59 protein complexes with 159 polypeptide chains | Large and strong interface hydrophobic patches | [139] |
| Lo Conte *et al.* | 1999 | 75 complexes (24 protease inhibitors, 19 antigen-antibody) 32 others (9 enzyme inhibitors, 11 signal transduction) | Observed interface non-polar characteristics as surface with few charged groups | [29] |
| Chakrabarti and Janin | 2002 | 70 complexes | Recognition patches in protein interfaces | [30] |
| Brinda *et al.* | 2002 | 20 homodimers | Graph-spectral analysis effectively identifies clusters at protein interfaces | [140] |
| Caffrey *et al* | 2004 | 64 complexes (42 homodimers, 12 heterodimers, 10 transient complexes) | Residue conservation is less than the core but more than the surface | [84] |
| Bahadur *et al.* | 2004 | 70 protein-protein complexes, 122 homodimers | Interface features (residue propensity score) in specific and non-specific complexes | [141] |
| Zhanhua *et al.* | 2005 | 156 heterodimers, 170 homodimers | High Hydrogen-bond density per interface residue with dominant charged and hydrophilic residues at the heterodimer protein interfaces | [69] |
| Pal *et al.* | 2007 | 204 protein complexes | Dominant peptide segments involved in specific interactions discriminate biological from non-biological ones | [90] |
| Reynolds *et al.* | 2009 | 220 heterodimers, 534 homodimers | PPI analysis server named ProtorP | [31] |
| Gromiha et al. | 2009 | 153 heterodimers | Dominance of aromatic and positively charged residues at interface | [142] |
| Guharoy and Chakrabarti | 2010 | 122 homodimer, 204 heterodimer | Conserved interface residues are not randomly distributed and are distinctly clustered | [91] |
| Sowmya *et al.* | 2011 | 192 heterodimer complexes | Heterodimeric interfaces are often abundant in polar residues | [7] |
| Chen *et al.* | 2013 | 113 heterodimer complexes | Direct correlation between binding affinity and amount of buried surface area at the interface | [143] |

standing the fundamentals of protein association. Thus, a combined formulation of structural and sequence features along with experimental evidence is often essential to improve the prediction accuracy of PPIs.

## Protein Interface Databases

Protein interactions have been extensively studied and deposited in various databases using interface datasets at protein and domain levels. PIBASE is a relational database of structurally defined interfaces obtained from protein domain pair interactions [95]. The database contains interfaces that are annotated based on geometric, physicochemical and topological properties responsible for the structural characterisation of protein complexes. Currently (as of February, 2013) contains 104,569 structures from PDB and PQS (Protein Quaternary Structure) and 598,638 domains from SCOP (Structural Classification of Proteins) and CATH (Class Architecture Topology and Homologous superfamily) databases with a total of 755,998 interfaces from the interfaces of SCOP [96], CATH [97] and interface chains. SCOPPI (Structural Classification of Protein-Protein Interfaces) database consists of classifications and annotations of protein domain interactions derived from PDB and SCOP domain definitions [98]. SCOPPI classifies domains based on their geometry and provides several interface characteristics such as number of interfaces, aminoacid types and positions, conservation, interface size, and permanent or transient nature of interactions. SCOPPI comprises of a total of 105,547 domain-domain interactions, 22,874 domain-domain interactions at 90% non-redundancy level and 4,630 family-family interactions with 15,058 interface types (February 2013).

The database of 3D Interacting Domain (3DID) archives domain-domain interactions obtained from high resolution protein structures [99]. 3DID provides information regarding structural similarity between different members of the same protein family. As of February 2013, 3DID contains 4302 unique Pfam domains participating in domain-domain interactions, with 175,144 domain-domain interactions of known 3D structures in 174,006 proteins. A large-scale protein domain interaction interface database, InterPare, is a public database and server for protein interface information obtained from 3D-structures [100]. The database detects interfaces by three methods i.e. calculating geometric distance method for checking distance between atoms in different domains, detecting Accessible Surface Area (ASA) and calculating Voronoi diagram which uses mathematical definition for interaction interfaces.

PDBsum [101, 102] is a pictorial database that provides summary information of 3D structures deposited at the PDB. The database provides structure diagrams, GO annotations, 1D sequence annotated by Pfam [103] and InterPro [104, 105] domain assignments, clefts in structures and schematic diagrams of PPI, including provisions for generating PDBsum information for user's own PDB file formats [102]. As of February 2013, PDBsum contains 91,642 entries including 1,854 superseded ones. Few other databases such as ProtCom (database of protein complexes) [106] and SCOWLP (Structural Characterisation of Water, Ligands and Protein) [107] provide protein interaction interface information derived from high-throughput data, experimentally determined structures at PDB and literature.

## Protein Interaction Characterisation and Prediction Tools/Servers

Several webservers and/or tools have been described over the past two decades for effective characterisation of protein interfaces based on physicochemical properties. InterPreTS developed by Aloy and Russell (2003) [108] predicts protein interactions in query sequences using similarity between protein sequences and known complexes. MODTIE predicts binary protein interaction based on similarity of query sequence with the structures at MODBASE [109]. SHARP2 server performs PPI prediction using patch analysis with parameters such as solvation potential, hydrophobicity, accessible surface area, interface residue propensity, planarity and protrusion. The Proface server described by Saha and colleagues dissects PPI to derive physical and chemical features [83]. Sppider [110] is a protein interface identification and recognition server based on solvent accessibility and structural information with high levels of accuracy (74%). The ProtorP server predicts physical and chemical characteristics of structural interfaces (types of atoms at interfaces, structural elements, H-bonds, bridging water molecules, salt bridges and disulphide bonds, interface area, planarity, eccentricity, gap volume and gap index *etc*. in homo and hetero complexes), to assess the different interface properties of the query protein against the pre-calculated set of interface features in different complexes [31]. Many other webservers developed based on structural data of available protein complexes at the PDB for the prediction of interaction sites and interfaces include cons-PPISP [111], meta-PPISP [112], PI2PE [113], PIC [114], LIGPLOT [115], ProTherm [116], PPI-Pred [117], MAPPIS [118], WHISCY [119], SNAPPI [120], PepSite [121], PrISE [122] and Wiki-pi [123]. Several other databases/ tools available for protein-protein interface hot-spots include ASEdb [124], BID [125], KFC [126], HotSprint [127], hotPOINT [128], APIS [129], PCRPI [130], HotRegion [131] and DBAC [132]. PPI hot spots and *in silico* techniques available for hot spot residues is reviewed by Morrow and Zhang [133]. The community wide experiment, CAPRI (Critical Assessment of Predicted Interactions) [134], performs continuous assessments of prediction models multiples times in a year. The accurate prediction of interaction regions and residue contacts while accommodating large conformational changes is often difficult [135]. The community experiment recently found that electrostatics and solvation terms marginally distinguish the designs of proteins from natural complexes, largely due to non-polar characteristics of interactions and also that the binding surfaces were structurally less embedded in designed monomers, suggesting the importance of conformational rigidity at the designed surface [136].

## CONCLUDING REMARKS

Molecular biology processes are frequently associated with interactions between protein pairs for specific functionality. PPIs have been widely studied using different approaches for an in-depth understanding of protein-protein recognition in cellular systems. Advances in the analyses of the interfaces give insights into the significance of prediction using sequence and structure information. The physical and chemical factors determining the interface formation is often multi-parametric in nature. Despite several line-of-thoughts in the area, current information lacks compelling reasons towards the formation of stable interface. This also hampers the incorporation of a comprehensive set of features to train predictors for reliable protein interaction predictions. Moreover, interfaces are part of surfaces in interacting monomers associated for specific functionality in biological units; mimicking these interface features under *in vivo* conditions poses further difficulty in prediction accuracy. Structural information is essential to obtain distinct features of binding for potential extrapolation to sequence level. Thus, understanding the basis of PPI using a dataset of known protein complexes and deriving important features responsible for binding is essential. Each individual protein complex portrays specific, selective and sensitive binding to its partner. An investigation on the combination of atomic features and residue types at the interface as compared to the surface in different classes of complexes is necessary to characterize binding sites. Therefore, learning the discriminative features of different PPIs from known complexes and consequently obtaining common patterns of recognition will be critical for predicting interacting partners.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

## ACKNOWLEDGEMENTS

Declared none.

## REFERENCES

[1] Reichmann, D.; Rahat, O.; Cohen, M.; Neuvirth, H.; Schreiber, G. The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.*, **2007**, *17*, 67-76.

[2] Robinson, C.V.; Sali, A.; Baumeister, W. The molecular sociology of the cell. *Nature*, **2007**, *450*, 973-82.

[3] Wells, J.A.; McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **2007**, *450*, 1001-9.

[4] Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.*, **2008**, *108*, 1225-44.

[5] Schreiber, G.; Keating, A.E. Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.*, **2011**, *21*, 50-61.

[6] Janin, J.; Bahadur, R.P.; Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q Rev. Biophys.*, **2008**, *41*, 133-80.

[7] Sowmya, G.; Anita, S.; Kangueane, P. Insights from the structural analysis of protein heterodimer interfaces. *Bioinformation*, **2011**, *6*, 137-43.

[8] Mrowka, R.; Patzak, A.; Herzel, H. Is there a bias in proteome research? *Genome Res.*, **2001**, *11*, 1971-3.

[9] von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S.G.; Fields, S.; Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **2002**, *417*, 399-403.

[10] Szilagyi, A.; Grimm, V.; Arakaki, A.K.; Skolnick, J. Prediction of physical protein-protein interactions. *Phys. Biol.*, **2005**, *2*, S1-16.

[11] Aloy, P.; Russell, R.B. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **2006**, *7*, 188-97.

[12] Huynen, M.A.; Snel, B.; von Mering, C.; Bork, P. Function prediction and protein networks. *Curr. Opin. Cell Biol.*, **2003**, *15*, 191-8.

[13] Smith, G.R.; Sternberg, M.J. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **2002**, *12*, 28-35.

[14] Wiehe, K.; Peterson, M.W.; Pierce, B.; Mintseris, J.; Weng, Z. Protein-protein docking: overview and performance analysis. *Methods Mol. Biol.*, **2008**, *413*, 283-314.

[15] Camacho, C.J.; Vajda, S. Protein-protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.*, **2002**, *12*, 36-40.

[16] Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, **2002**, *47*, 409-43.

[17] Heifetz, A.; Katchalski-Katzir, E.; Eisenstein, M. Electrostatics in protein-protein docking. *Protein Sci.*, **2002**, *11*, 571-87.

[18] Brooijmans, N.; Kuntz, I.D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, **2003**, *32*, 335-73.

[19] Kastritis, P.L.; Bonvin, A.M. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.*, **2010**, *9*, 2216-25.

[20] Lensink, M.F.; Mendez, R.; Wodak, S.J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*, **2007**, *69*, 704-18.

[21] Dobbins, S.E.; Lesk, V.I.; Sternberg, M.J. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. USA*, **2008**, *105*, 10390-5.

[22] Gunasekaran, K.; Nussinov, R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.*, **2007**, *365*, 257-73.

[23] Huang, B.; Schroeder, M. Using protein binding site prediction to improve protein docking. *Gene*, **2008**, *422*, 14-21.

[24] Zhang, Q.; Sanner, M.; Olson, A.J. Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins*, **2009**, *75*, 453-67.

[25] Janin, J.; Chothia, C. The structure of protein-protein recognition sites. *J. Biol. Chem.*, **1990**, *265*, 16027-30.

[26] Jones, S.; Thornton, J.M. Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, **1995**, *63*, 31-65.

[27] Jones, S.; Thornton, J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, **1996**, *93*, 13-20.

[28] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.*, **2000**, *28*, 235-42.

[29] Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **1999**, *285*, 2177-98.

[30] Chakrabarti, P.; Janin, J. Dissecting protein-protein recognition sites. *Proteins*, **2002**, *47*, 334-43.

[31] Reynolds, C.; Damerell, D.; Jones, S. ProtorP: a protein-protein interaction analysis server. *Bioinformatics*, **2009**, *25*, 413-4.

[32] Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y. L.; Ooi, C. E.; Godwin, B.; Vitols, E.; Vijayadamodar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C. A.; Finley, R. L., Jr.; White, K. P.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, R. A.; McKenna, M. P.; Chant, J.; Rothberg, J. M. A protein interaction map of Drosophila melanogaster. *Science*, **2003**, *302*, 1727-36.

[33] Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P. O.; Han, J. D.; Chesneau, A.; Hao, T.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J. F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. A map of the interactome network of the metazoan C. elegans. *Science*, **2004**, *303*, 540-3.

[34] Ozbabacan, S. E.; Engin, H. B.; Gursoy, A.; Keskin, O. Transient protein-protein interactions. *Protein Eng Des Sel*, **2011**, *24*, 635-48.

[35] Aloy, P.; Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, **2004**, *22*, 1317-21.

[35] Aloy, P.; Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **2004**, *22*, 1317-21.

[36] Hart, G. T.; Ramani, A. K.; Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **2006**, *7*, 120.

[37] Sambourg, L.; Thierry-Mieg, N. New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics*, **2010**, *11*, 605.

[38] Piehler, J. New methodologies for measuring protein interactions in vivo and in vitro. *Curr. Opin. Struct. Biol.*, **2005**, *15*, 4-14.

[39] Fields, S.; Song, O. A novel genetic system to detect protein-protein interactions. *Nature*, **1989**, *340*, 245-6.

[40] Vojtek, A. B.; Hollenberg, S. M.; Cooper, J. A. Mammalian Ras interacts directly with the serine/threonine kinase Raf. *Cell*, **1993**, *74*, 205-14.

[41] Golemis, E. A.; Khazak, V. Alternative yeast two-hybrid systems. The interaction trap and interaction mating. *Methods Mol. Biol.*, **1997**, *63*, 197-218.

[42] Aronheim, A. Improved efficiency sos recruitment system: expression of the mammalian GAP reduces isolation of Ras GTPase false positives. *Nucleic Acids Res.*, **1997**, *25*, 3373-4.

[43] Wehr, M. C.; Laage, R.; Bolz, U.; Fischer, T. M.; Grunewald, S.; Scheek, S.; Bach, A.; Nave, K. A.; Rossner, M. J. Monitoring regulated protein-protein interactions using split TEV. *Nat. Methods*, **2006**, *3*, 985-93.

[44] Karimova, G.; Pidoux, J.; Ullmann, A.; Ladant, D. A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 5752-6.

[45] Joung, J. K.; Ramm, E. I.; Pabo, C. O. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 7382-7.

[46] Eyckerman, S.; Verhee, A.; der Heyden, J. V.; Lemmens, I.; Ostade, X. V.; Vandekerckhove, J.; Tavernier, J. Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.*, **2001**, *3*, 1114-9.

[47] Hu, C. D.; Kerppola, T. K. Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis. *Nat. Biotechnol.*, **2003**, *21*, 539-45.

[48] Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **2002**, *415*, 141-7.

[49] Borch, J.; Roepstorff, P.; Moller-Jensen, J. Nanodisc-based co-immunoprecipitation for mass spectrometric identification of membrane-interacting proteins. *Mol. Cell Proteomics*, **2011**, *10*, O110 006775.

[50] Moresco, J. J.; Carvalho, P. C.; Yates, J. R., 3rd. Identifying components of protein complexes in C. elegans using co-immunoprecipitation and mass spectrometry. *J. Proteomics*, **2010**, *73*, 2198-204.

[51] Cayenne, A. P.; Gabert, B.; Stillman, J. H. Identification of proteins interacting with lactate dehydrogenase in claw muscle of the porcelain crab Petrolisthes cinctipes. *Comp. Biochem. Physiol. Part D Genomics Proteomics*, **2011**, *6*, 393-8.

[52] Shoemaker, B. A.; Panchenko, A. R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **2007**, *3*, e42.

[53] Tuncbag, N.; Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform.*, **2009**, *10*, 217-32.

[54] Xenarios, I.; Salwinski, L.; Duan, X. J.; Higney, P.; Kim, S. M.; Eisenberg, D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **2002**, *30*, 303-5.

[55] Duan, X. J.; Xenarios, I.; Eisenberg, D. Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. *Mol. Cell Proteomics*, **2002**, *1*, 104-16.

[56] Bowers, P. M.; Pellegrini, M.; Thompson, M. J.; Fierro, J.; Yeates, T. O.; Eisenberg, D. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **2004**, *5*, R35.

[57] Bader, G. D.; Hogue, C. W. BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, **2000**, *16*, 465-77.

[58] Alfarano, C.; Andrade, C. E.; Anthony, K.; Bahroos, N.; Bajec, M.; Bantoft, K.; Betel, D.; Bobechko, B.; Boutilier, K.; Burgess, E.; Buzadzija, K.; Cavero, R.; D'Abreo, C.; Donaldson, I.; Dorairajoo, D.; Dumontier, M. J.; Dumontier, M. R.; Earles, V.; Farrall, R.; Feldman, H.; Garderman, E.; Gong, Y.; Gonzaga, R.; Grytsan, V.; Gryz, E.; Gu, V.; Haldorsen, E.; Halupa, A.; Haw, R.; Hrvojic, A.; Hurrell, L.; Isserlin, R.; Jack, F.; Juma, F.; Khan, A.; Kon, T.; Konopinsky, S.; Le, V.; Lee, E.; Ling, S.; Magidin, M.; Moniakis, J.; Montojo, J.; Moore, S.; Muskat, B.; Ng, I.; Paraiso, J. P.; Parker, B.; Pintilie, G.; Pirone, R.; Salama, J. J.; Sgro, S.; Shan, T.; Shu, Y.; Siew, J.; Skinner, D.; Snyder, K.; Stasiuk, R.; Strumpf, D.; Tuekam, B.; Tao, S.; Wang, Z.; White, M.; Willis, R.; Wolting, C.; Wong, S.; Wrong, A.; Xin, C.; Yao, R.; Yates, B.; Zhang, S.; Zheng, K.; Pawson, T.; Ouellette, B. F.; Hogue, C. W. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **2005**, *33*, D418-24.

[59] Isserlin, R.; El-Badrawi, R. A.; Bader, G. D. The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)*, **2011**, *2011*, baq037.

[60] Chatr-aryamontri, A.; Ceol, A.; Palazzi, L. M.; Nardelli, G.; Schneider, M. V.; Castagnoli, L.; Cesareni, G. MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **2007**, *35*, D572-4.

[61] Kerrien, S.; Aranda, B.; Breuza, L.; Bridge, A.; Broackes-Carter, F.; Chen, C.; Duesbury, M.; Dumousseau, M.; Feuermann, M.; Hinz, U.; Jandrasits, C.; Jimenez, R. C.; Khadake, J.; Mahadevan, U.; Masson, P.; Pedruzzi, I.; Pfeiffenberger, E.; Porras, P.; Raghunath, A.; Roechert, B.; Orchard, S.; Hermjakob, H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **2012**, *40*, D841-6.

[62] Breitkreutz, B. J.; Stark, C.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Livstone, M.; Oughtred, R.; Lackner, D. H.; Bahler, J.; Wood, V.; Dolinski, K.; Tyers, M. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **2008**, *36*, D637-40.

[63] Keshava Prasad, T. S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; Balakrishnan, L.; Marimuthu, A.; Banerjee, S.; Somanathan, D. S.; Sebastian, A.; Rani, S.; Ray, S.; Harrys Kishore, C. J.; Kanth, S.; Ahmed, M.; Kashyap, M. K.; Mohmood, R.; Ramachandra, Y. L.; Krishna, V.; Rahiman, B. A.; Mohan, S.; Ranganathan, P.; Ramabadran, S.; Chaerkady, R.; Pandey, A. Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, **2009**, *37*, D767-72.

[64] Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C.; Jensen, L. J. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **2013**, *41*, D808-15.

[65] Nooren, I. M.; Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.*, **2003**, *22*, 3486-92.

[66] Karthikraja, V.; Suresh, A.; Lulu, S.; Kangueane, U.; Kangueane, P. Types of interfaces for homodimer folding and binding. *Bioinformation*, **2009**, *4*, 101-11.

[67] Suresh, A.; Karthikraja, V.; Lulu, S.; Kangueane, U.; Kangueane, P. A decision tree model for the prediction of homodimer folding mechanism. *Bioinformation*, **2009**, *4*, 197-205.

[68] Jones, S.; Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **1997**, *272*, 121-32.

[69] Zhanhua, C.; Gan, J. G.; Lei, L.; Sakharkar, M. K.; Kangueane, P. Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation*, **2005**, *1*, 28-39.

[70] Zhang, Z.; Witham, S.; Alexov, E. On the role of electrostatics in protein-protein interactions. *Phys. Biol.*, **2011**, *8*, 035001.

[71] Nooren, I. M.; Thornton, J. M. Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.*, **2003**, *325*, 991-1018.

[72] La, D.; Kong, M.; Hoffman, W.; Choi, Y. I.; Kihara, D. Predicting permanent and transient protein-protein interfaces. *Proteins*, **2012**.

[73] Ofran, Y.; Rost, B. Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **2003**, *325*, 377-87.

[74] Ezkurdia, I.; Bartoli, L.; Fariselli, P.; Casadio, R.; Valencia, A.; Tress, M. L. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform.*, **2009**, *10*, 233-46.

[75] Perkins, J. R.; Diboun, I.; Dessailly, B. H.; Lees, J. G.; Orengo, C. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **2010**, *18*, 1233-43.

[76] Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins*, **2010**, *78*, 3111-4.

[77] Swapna, L. S.; Bhaskara, R. M.; Sharma, J.; Srinivasan, N. Roles of residues in the interface of transient protein-protein complexes before complexation. *Sci. Rep.*, **2012**, *2*, 334.

[78] Jones, S.; Marin, A.; Thornton, J. M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, **2000**, *13*, 77-82.

[79] Park, S. H.; Reyes, J. A.; Gilbert, D. R.; Kim, J. W.; Kim, S. Prediction of protein-protein interaction types using association rule based classification. *BMC Bioinformatics*, **2009**, *10*, 36.

[80] Chothia, C.; Janin, J. Principles of protein-protein recognition. *Nature*, **1975**, *256*, 705-8.

[81] Miller, S.; Lesk, A. M.; Janin, J.; Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature*, **1987**, *328*, 834-6.

[82] Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.*, **1997**, *6*, 53-64.

[83] Saha, R. P.; Bahadur, R. P.; Pal, A.; Mandal, S.; Chakrabarti, P. ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Struct. Biol.*, **2006**, *6*, 11.

[84] Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **2004**, *13*, 190-202.

[85] Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **2003**, *53*, 708-19.

[86] Guharoy, M.; Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 15447-52.

[87] Ofran, Y.; Rost, B. ISIS: interaction sites identified from sequence. *Bioinformatics*, **2007**, *23*, e13-6.

[88] Xu, D.; Tsai, C. J.; Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.*, **1997**, *10*, 999-1012.

[89] Gao, Y.; Wang, R.; Lai, L. Structure-based method for analyzing protein-protein interfaces. *J. Mol. Model.*, **2004**, *10*, 44-54.

[90] Pal, A.; Chakrabarti, P.; Bahadur, R.; Rodier, F.; Janin, J. Peptide segments in protein-protein interfaces. *J. Biosci.*, **2007**, *32*, 101-11.

[91] Guharoy, M.; Chakrabarti, P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics*, **2010**, *11*, 286.

[92] Planas-Iglesias, J.; Bonet, J.; Garcia-Garcia, J.; Marin-Lopez, M. A.; Feliu, E.; Oliva, B. Understanding protein-protein interactions using local structural features. *J. Mol. Biol.*, **2013**, *425*, 1210-24.

[93] Gromiha, M. M.; Saranya, N.; Selvaraj, S.; Jayaram, B.; Fukui, K. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome Sci.*, **2011**, *9 Suppl 1*, S13.

[94] Mosca, R.; Ceol, A.; Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods*, **2012**, *10*, 47-53.

[95] Davis, F. P.; Sali, A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **2005**, *21*, 1901-7.

[96] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **1995**, *247*, 536-40.

[97] Knudsen, M.; Wiuf, C. The CATH database. *Hum. Genomics*, **2010**, *4*, 207-12.

[98] Winter, C.; Henschel, A.; Kim, W. K.; Schroeder, M. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **2006**, *34*, D310-4.

[99] Stein, A.; Russell, R. B.; Aloy, P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **2005**, *33*, D413-7.

[100] Gong, S.; Park, C.; Choi, H.; Ko, J.; Jang, I.; Lee, J.; Bolser, D. M.; Oh, D.; Kim, D. S.; Bhak, J. A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **2005**, *6*, 207.

[101] Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **2001**, *29*, 221-2.

[102] Laskowski, R. A. PDBsum new things. *Nucleic Acids Res.*, **2009**, *37*, D355-9.

[103] Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. The Pfam protein families database. *Nucleic Acids Res.*, **2004**, *32*, D138-41.

[104] Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Birney, E.; Biswas, M.; Bucher, P.; Cerutti, L.; Corpet, F.; Croning, M. D.; Durbin, R.; Falquet, L.; Fleischmann, W.; Gouzy, J.; Hermjakob, H.; Hulo, N.; Jonassen, I.; Kahn, D.; Kanapin, A.; Karavidopoulou, Y.; Lopez, R.; Marx, B.; Mulder, N. J.; Oinn, T. M.; Pagni, M.; Servant, F.; Sigrist, C. J.; Zdobnov, E. M. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **2001**, *29*, 37-40.

[105] Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Buillard, V.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Daugherty, L.; Dibley, M.; Finn, R.; Fleischmann, W.; Gough, J.; Haft, D.; Hulo, N.; Hunter, S.; Kahn, D.; Kanapin, A.; Kejariwal, A.; Labarga, A.; Langendijk-Genevaux, P. S.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Orengo, C.; Petryszak, R.; Selengut, J. D.; Sigrist, C. J.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H.; Yeats, C. New developments in the InterPro database. *Nucleic Acids Res.*, **2007**, *35*, D224-8.

[106] Kundrotas, P. J.; Alexov, E. PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res.*, **2007**, *35*, D575-9.

[107] Teyra, J.; Doms, A.; Schroeder, M.; Pisabarro, M. T. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **2006**, *7*, 104.

[108] Aloy, P.; Russell, R. B. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **2003**, *19*, 161-2.

[109] Pieper, U.; Eswar, N.; Davis, F. P.; Braberg, H.; Madhusudhan, M. S.; Rossi, A.; Marti-Renom, M.; Karchin, R.; Webb, B. M.; Eramian, D.; Shen, M. Y.; Kelly, L.; Melo, F.; Sali, A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **2006**, *34*, D291-5.

[110] Porollo, A.; Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins*, **2007**, *66*, 630-45.

[111] Chen, H.; Zhou, H. X. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins*, **2005**, *61*, 21-35.

[112] Qin, S.; Zhou, H. X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, **2007**, *23*, 3386-7.

[113] Tjong, H.; Qin, S.; Zhou, H. X. PI2PE: protein interface/interior prediction engine. *Nucleic Acids Res.*, **2007**, *35*, W357-62.

[114] Tina, K. G.; Bhadra, R.; Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.*, **2007**, *35*, W473-6.

[115] Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **1995**, *8*, 127-34.

[116] Gromiha, M. M.; An, J.; Kono, H.; Oobatake, M.; Uedaira, H.; Prabakaran, P.; Sarai, A. ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **2000**, *28*, 283-5.

[117] Bradford, J. R.; Westhead, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **2005**, *21*, 1487-94.

[118] Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.*, **2007**, *5*, 43.

[119] de Vries, S. J.; van Dijk, A. D.; Bonvin, A. M. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins*, **2006**, *63*, 479-89.

[120] Jefferson, E. R.; Walsh, T. P.; Roberts, T. J.; Barton, G. J. SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. *Nucleic Acids Res.*, **2007**, *35*, D580-9.

[121] Trabuco, L. G.; Lise, S.; Petsalaki, E.; Russell, R. B. PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Res.*, **2012**, *40*, W423-7.

[122] Jordan, R. A.; El-Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinformatics*, **2012**, *13*, 41.

[123] Orii, N.; Ganapathiraju, M. K. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PLoS One*, **2012**, *7*, e49029.

[124] Thorn, K. S.; Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **2001**, *17*, 284-5.

[125] Fischer, T. B.; Arunachalam, K. V.; Bailey, D.; Mangual, V.; Bakhru, S.; Russo, R.; Huang, D.; Paczkowski, M.; Lalchandani, V.; Ramachandra, C.; Ellison, B.; Galer, S.; Shapley, J.; Fuentes, E.; Tsai, J. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **2003**, *19*, 1453-4.

[126] Darnell, S. J.; LeGault, L.; Mitchell, J. C. KFC Server: interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.*, **2008**, *36*, W265-9.

[127] Guney, E.; Tuncbag, N.; Keskin, O.; Gursoy, A. HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **2008**, *36*, D662-6.

[128] Tuncbag, N.; Keskin, O.; Gursoy, A. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **2010**, *38*, W402-6.

[129] Xia, J. F.; Zhao, X. M.; Song, J.; Huang, D. S. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **2010**, *11*, 174.

[130] Assi, S. A.; Tanaka, T.; Rabbitts, T. H.; Fernandez-Fuentes, N. PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res.*, **2010**, *38*, e86.

[131] Cukuroglu, E.; Gursoy, A.; Keskin, O. HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res.*, **2012**, *40*, D829-33.

[132] Li, Z.; Wong, L.; Li, J. DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts. *BMC Syst. Biol.*, **2011**, *5 Suppl 1*, S5.

[133] Morrow, J. K.; Zhang, S. Computational prediction of protein hot spot residues. *Curr. Pharm. Des.*, **2012**, *18*, 1255-65.

[134] Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, **2003**, *52*, 2-9.

[135] Janin, J. Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci.*, **2005**, *14*, 278-83.

[136] Fleishman, S. J.; Whitehead, T. A.; Strauch, E. M.; Corn, J. E.; Qin, S.; Zhou, H. X.; Mitchell, J. C.; Demerdash, O. N.; Takeda-Shitaka, M.; Terashi, G.; Moal, I. H.; Li, X.; Bates, P. A.; Zacharias, M.; Park, H.; Ko, J. S.; Lee, H.; Seok, C.; Bourquard, T.; Bernauer, J.; Poupon, A.; Aze, J.; Soner, S.; Ovali, S. K.; Ozbek, P.; Tal, N. B.; Haliloglu, T.; Hwang, H.; Vreven, T.; Pierce, B. G.; Weng, Z.; Perez-Cano, L.; Pons, C.; Fernandez-Recio, J.; Jiang, F.; Yang, F.; Gong, X.; Cao, L.; Xu, X.; Liu, B.; Wang, P.; Li, C.; Wang, C.; Robert, C. H.; Guharoy, M.; Liu, S.; Huang, Y.; Li, L.; Guo, D.; Chen, Y.; Xiao, Y.; London, N.; Itzhaki, Z.; Schueler-Furman, O.; Inbar, Y.; Potapov, V.; Cohen, M.; Schreiber, G.; Tsuchiya, Y.; Kanamori, E.; Standley, D. M.; Nakamura, H.; Kinoshita, K.; Driggers, C. M.; Hall, R. G.; Morgan, J. L.; Hsu, V. L.; Zhan, J.; Yang, Y.; Zhou, Y.; Kastritis, P. L.; Bonvin, A. M.; Zhang, W.; Camacho, C. J.; Kilambi, K. P.; Sircar, A.; Gray, J. J.; Ohue, M.; Uchikoga, N.; Matsuzaki, Y.; Ishida, T.; Akiyama, Y.; Khashan, R.; Bush, S.; Fouches, D.; Tropsha, A.; Esquivel-Rodriguez, J.; Kihara, D.; Stranges, P. B.; Jacak, R.; Kuhlman, B.; Huang, S. Y.; Zou, X.; Wodak, S. J.; Janin, J.; Baker, D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J. Mol. Biol.*, **2011**, *414*, 289-302.

[137] Chothia, C.; Wodak, S.; Janin, J. Role of subunit interfaces in the allosteric mechanism of hemoglobin. *Proc. Natl. Acad. Sci. USA*, **1976**, *73*, 3793-7.

[138] Dasgupta, S.; Iyer, G. H.; Bryant, S. H.; Lawrence, C. E.; Bell, J. A. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, **1997**, *28*, 494-514.

[139] Lijnzaad, P.; Argos, P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, **1997**, *28*, 333-43.

[140] Brinda, K. V.; Kannan, N.; Vishveshwara, S. Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng.*, **2002**, *15*, 265-77.

[141] Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **2004**, *336*, 943-55.

[142] Gromiha, M. M.; Yokota, K.; Fukui, K. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Mol. Biosyst.*, **2009**, *5*, 1779-86.

[143] Chen, J.; Sawyer, N.; Regan, L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.*, **2013**, *22*, 510-5.

**Received: March 4, 2013** **Revised: April 12, 2013** **Accepted: April 15, 2013**

# Chapter 2: Methods and applications

Methods and applications that were developed and used in this study are summarised in Table 2.1. The ensuing publications have also been listed and included in the relevant chapter.

**Table 2.1: Methods, applications and publications**

| Methods/Applications | Chapter | Thesis Publication |
|---|---|---|
| Protein interface residue-level classes and their discriminatory structural features | 3 | 2 |
| Protein interfaces and biological functions | 4 | 3 |
| Dissecting interfaces of interacting proteins: integrin αvβ6·uPAR interactions | 5 | 4 |

# Chapter 3: Protein interface residue-level classes and their discriminatory structural features

## 3.1 Summary

PPI establishes the central basis for complex cellular networks in a biological cell. Investigation on protein interfaces of known complexes is an important step towards deciphering the driving forces of PPIs. Each PPI complex is specific, sensitive and selective to binding. Therefore, an investigation on the relative interface-surface polarity of each complex is essential to determine the predominant forces driving binding.

In this study, a comprehensive structural analysis of 278 non-redundant heterodimeric protein complexes from the PDB has been carried out. The relative surface-interface polarities, referred to as interface polarity abundance of each complex in the dataset were estimated for predominance of polar and/or non-polar interactions at the protein interface. This property divides the dataset into two interface classes as also observed in our previous study with a smaller dataset [24]. The complexes with less polar residues at the interface as compared to the surface (~60%), which is the 'classical' definition of a PPI complex, are designated as 'class A' (with predominant non-polar interactions at the interface), while the complexes with more polar residues at the interface as compared to the surface (~40%), are designated as 'class B' (with predominant polar interactions at the interface).

The essential PPI structural features such as interface area (ΔASA), the relative abundance of polar and non-polar residues at the interface (interface property abundance), hydrogen bonds (H-bonds), salt bridges, percentage of charged residues at the interface (interface charged residues%), solvation free energy gain upon interface formation ($\Delta^{i}G$), binding energy (BE), and electrostatics among these interface classes were investigated and their gleaned features are documented. Water molecules and ions are not present in all the complexes in this dataset and their role has therefore not been explicitly considered in this analysis. The need for a residue-level characterization of the interface in addition to other structural features is discussed (Publication 2).

## 3.2 Publication 2

# Discrete structural features among interface residue-level classes

**Gopichandran Sowmya[1], Shoba Ranganathan[1],***

[1]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109 Australia

*Corresponding author

Email addresses:

GS: sowmya.gopichandran@mq.edu.au

SR: shoba.ranganathan@mq.edu.au

**Abbreviations**

PPI, Protein-Protein Interaction; $\Delta^{i}G$, solvent free energy gain upon interface formation; BE, Binding Energy; H-bonds, Hydrogen bonds; PDB, Protein Data Bank; ΔASA, interface area; P, Polar residues; NP, Non-polar residues; S, Surface polarity; I, Interface Polarity.

**Keywords:** Protein-protein interaction (PPI), heterodimers, interface, surface, polarity.

# Abstract

### Background

Protein-protein interaction (PPI) is essential for molecular functions in biological cells. Investigation on protein interfaces of known complexes is an important step towards deciphering the driving forces of PPIs. Each PPI complex is specific, sensitive and selective to binding. Therefore, we have estimated the relative difference in percentage of polar residues between surface and the interface for each complex in a non-redundant heterodimer dataset of 278 complexes to understand the predominant forces driving binding.

### Results

Our analysis showed ~60% of protein complexes with surface polarity greater than interface polarity (designated as class A). However, a considerable number of complexes (~40%) have interface polarity greater than surface polarity, (designated as class B), with a significantly different p-value of 1.66E-45 from class A. Comprehensive analyses of protein complexes show that interface features such as interface area, interface polarity abundance, solvation free energy gain upon interface formation, binding energy and the percentage of interface charged residue abundance distinguish among class A and class B complexes, while electrostatic visualization maps also help differentiate interface classes among complexes.

### Conclusions

Class A complexes are classical with abundant non-polar interactions at the interface; however class B complexes have abundant polar interactions at the interface, similar to protein surface characteristics. Five physicochemical interface features analyzed from the protein heterodimer dataset are discriminatory among the interface residue-level classes. These novel observations find application in developing residue-level models for protein-protein binding prediction, protein-protein docking studies and interface inhibitor design as drugs.

## Background

Protein-protein binding is a known phenomenon in complex biological networks. The molecular principle of such binding is often elusive in nature. Understanding its driving forces using known protein complexes is essential. The analysis of existing protein-protein interaction (PPI) complexes from the Protein Data Bank (PDB) [1] is the key to gaining insights into recognition mechanisms and binding principles as reviewed elsewhere [2-6]. Sequence and structural investigations on the existing complexes has been carried out for several decades [3, 7-10]. In these extensive surveys, structural features over diverse datasets of protein-protein complexes were typically averaged, obscuring information on individual proteins' structural integrity. Each individual complex is specific and sensitive to binding. Although, non-polar (or hydrophobic) interactions are known to play a major role in contributing to the driving force for binding, in a considerable number of complexes, polar interactions are also observed to contribute abundantly to the formation of a stable interface [11]. Therefore, it is often essential to study the relative difference in surface and interface polarity of each PPI complex to determine the major binding forces at the interface and determine their discriminatory features.

Interfaces are localized regions of surfaces with different physico-chemical properties compared to the rest of the surfaces, thereby driving binding to other molecules. Both physical and chemical features (including hydrophobicity, electrostatic interactions, binding energy, interface size, hydrogen bonds, salt bridges, disulphide bonds, planarity, sphericity, shape complementarity, amino acid chemical groups, and conserved residue clusters) govern the formation of protein interfaces as described elsewhere [7, 9, 12-18]. The chemical nature of residues forming a protein interface (amino acid residue composition) determines the hydrophobic effect of an interface. Non-polar (or hydrophobic) residues are observed to occur predominantly at the protein interface, playing a major role in contributing to the driving force for binding [7, 13]. Interfaces are observed to be less non-polar (or hydrophobic) than the protein interior [13]. The residue composition of protein-protein interfaces was observed to be more similar to the protein surface than the protein interior [9].

Interfaces were observed to be significantly polar as well as non-polar with few charged groups, similar to the characteristics of the protein surface [12]. Structural analysis also revealed that charged and polar amino acids are involved in protein-protein interactions as reviewed elsewhere [19]. Charged and polar residues contributing to binding specificity and complex formation are demonstrated in a number of complexes such as human IL-4, human CD2 and CD58, barnase-barstar, Colicin E9, integrin αvβ6 membrane protein and in intrinsically disordered proteins [20-25]. Shape complementarity, polar interactions, hydrogen bonding and salt bridges are also known to contribute to binding specificity and free energy of binding [17, 24, 26, 27]. Also, charged and aromatic side chains are crucial for binding, determining the cation-pi, electrostatic and aromatic interactions [8]. The role of electrostatics in binding stability of protein-protein complexes is demonstrated [16]. These observations indicate that although PPIs are driven by non-polar interactions at the interface for a majority of complexes, in some cases polar interactions contribute to binding specificity (characteristic of polar residues) and likewise to complex stabilization. Therefore, a study on the relative percentage difference between surface and interface polarities of each protein complex is often essential. In our previous study, we have identified a class of complexes (class B) with more polar residues that core and surface, where binding is mainly polar with a dataset of 198 complexes [11]. This observation has now been extended for an updated yet non-redundant dataset of 278 protein complexes to verify and identify any discriminatory features among these interface residue-level classes.

In this study, we have carried out a comprehensive structural analysis of 278 non-redundant heterodimeric protein complexes from the PDB. We estimated the relative difference in surface and interface polarities of each complex in the dataset, using percentage values of polar residues. This property divides the dataset into two interface classes as also observed in our previous study with a smaller dataset [11]. Class A has less polar residues at the interface than the rest of the surface (~60%) which is the 'classical' definition of a PPI complex and class B has more polar residues at the interface than the rest of the surface (~40%), is 'non-classical.' Therefore, we have investigated essential PPI structural features such as interface area (ΔASA), the relative abundance of polar and non-polar residues at the interface (interface polarity abundance), hydrogen bonds (H-bonds), salt bridges, percentage of charged residues at the interface (interface charged residues%), solvation free energy gain upon interface formation ($\Delta^{i}G$), binding energy (BE), and electrostatics among these interface classes and their gleaned features are documented. We identified five key features (ΔASA, interface polarity abundance, interface charged residues%, $\Delta^{i}G$ and BE) that are significantly different between the interface classes. These novel observations have implications for residue-level characterization of protein complexes to develop models for protein-protein binding prediction and docking studies.

# Methods

### Heterodimer dataset

We created a non-redundant heterodimer dataset of protein complexes from the PDB, using the RCSB PDB's advanced search interface. The following criteria were used for filtering: (i) resolution <= 3Å (ii) protein size >50 residues (iii) contains experimental data (iv) number of chains, entities and oligomeric state is set at 2 (v) devoid of DNA or RNA or a hybrid of such molecules with the protein or otherwise (vi) entries with mutations were not included and (vii) sequence identity cut-off is set to 30%, which is the minimum cut-off available in the PDB. As a second step, the USEARCH program [28] was used to further remove the redundancy among heterodimer complexes at sequence identity cut-off of 20%, as this threshold eliminates remote homology up to 25% sequence identity seen in structures as well [29].

### Interface analysis

The interface of PPI complex is calculated as the change in solvent accessible surface area (ΔASA) upon complex formation. The Surface Racer 5.0 program [30] was used to calculate ASA with a probe radius of 1.4Å and Lee and Richards implementation [31]. Interface residues with ΔASA > 0.1Å$^2$ were considered for this analysis, as defined by Chakrabarti and Janin [32]. ASA was used to determine surface residues of each complex. The amount of polar, non-polar and charged residues at the interface was then estimated for the dataset. The interface polarity abundance (P%-NP%) is measured as the difference in the percentage of polar residues (P%) and percentage of non-polar residues (NP%) at the interface [11].

### Classification based on relative interface-surface polarity

Interfaces are part of protein surface formed upon binding of individual subunits. Each protein complex has a specific composition of polar (P) and nonpolar (NP) residues at the surface (S) and at the interface (I). The distribution of polar and nonpolar residues at the interface of a protein complex describes the nature of the interface and the major driving force for binding. We have calculated the percentage of polar and nonpolar residues at the surface and interface for each complex in the dataset. Polar residues considered in the analysis are R, N, D, E, Q, H, K, S, T, and Y and non-polar residues are A, C, G, I, L, M, F, P, V, and W. Complexes were then grouped based on the relative difference in percentage of polar residues between surface (S) and the interface (I). Complexes with interface polarity less than the surface (represented as S>I) are grouped as class A, and those that have interface polarity greater than the surface (represented as S<I) are grouped as class B [11].

### Intermolecular H-bonds and salt bridges calculation

We calculated the intermolecular hydrogen bonds for the dataset using HBPLUS program [33] at a distance of < 4Å. The H-bonds were extracted such that the donor and acceptor are from two different chains. Salt bridges were calculated using SBION program [34] within a distance of 4Å. The salt bridges were also extracted such that the oppositely charged atoms are from two different chains.

### $\Delta^i$G and BE calculation

PDBePISA webserver [35] was used to obtain the solvation free energy gain upon interface formation ($\Delta^i$G, in kcal/mol, with negative $\Delta^i$G values indicating hydrophobic interface) and for the heterodimer dataset. BE values were calculated using the DCOMPLEX program [36] with the most negative value considered the strongest. The DCOMPLEX program uses DFIRE-based potentials [37] to calculate BE terms, without values for individual components (electrostatic, van der Waals, hydrophobic and entropic terms) contributing to BE. Initially, the program calculates the total atom-atom potential of mean force, G, for each protein structure as follows:

$$G = \frac{1}{2} \sum_{i,j} \bar{u}(_{i,j}, r_{i,j}) \qquad (1)$$

where ū is the atom-atom potential of mean force between two atoms, i and j which are ‘r’ distance apart. The total is over atomic pairs which are not from the same residue and a K factor is used to avoid double-counting of residue-residue and atom-atom interactions [36]. The binding energy between two interacting proteins A and B can be calculated as follows:

$$\Delta G_{bind} = G_{complex}(G_A + G_B) \qquad (2)$$

where A and B are considered as two protein structures whose interface residues contribute most to $\Delta G_{bind}$. Therefore, DCOMPLEX [36] uses the equation below to calculate BE:

$$\Delta G_{\mathrm{bind}} = \frac{1}{2} \sum_{ij}^{interface} \bar{u}(_{i,j}, r_{i,j}) \qquad (3)$$

### Electrostatic potential at the interface

The surface electrostatic potential of chain A and chain B of a protein complex was calculated by solving Poisson-Boltzmann equation with dielectric constant (protein) of 4 using DEEPVIEW [38].

### Statistical analysis

The Wilcoxon signed-rank test [39], a non-parametric statistical hypothesis test is used to compare the two interface classes to assess whether the mean ranks for the PPI features in the two classes differ (i.e. it is a paired difference test). The discriminatory PPI features among the two classes were thus tested for statistical significance with $p < 0.05$ (for the Wilcoxon signed-rank test) in RStudio [40].

## Results and Discussion

We calculated the amount of polar and non-polar residues at the surface and interface of each protein-protein complex and estimated their relative interface-surface polarities for classification into class A and class B (as described in Materials and Methods section), to determine the type of interactions predominantly driving protein-protein binding. Additional File 1: **Table S1** shows the heterodimer dataset (278) divided into class A (165) and class B (113). Thus, 59.4% of complexes in our dataset belong to class A (relative surface polarity is greater than interface polarity), where non-polar interactions are predominant at the interface, as previously observed in a number of studies [7, 13]. Nevertheless, 40.6% of complexes belong to class B (relative interface polarity is greater than surface polarity), where polar interactions are predominant at the interface, similar to the surface characteristics as also observed [12]. Class A and class B are significantly different with a p-value of 1.66E-45 (using Wilcoxon rank sum test) as shown in Additional File 2: **Figure S1**. Examples of class A and class B complexes representing predominant non-polar and predominant polar interfaces (using the PDBsum [41] interaction analysis) respectively are shown in **Figure 1**.

### PPI features among class A and class B complexes

We carried out a statistical analysis of all the structural features (described in Materials and Methods section including ΔASA, interface polarity abundance, interface charged residues%, H-bonds, salt bridges, $\Delta^i G$, BE) in R program (using Wilcoxon rank sum test), to determine whether structural features discriminate among class A and class B complexes. Interestingly, five structural features showed significant difference among the interface classes as shown in **Figure 2**, with p-value < 0.05 (**Table 1**). The q-value in Table 1 is the smallest False Discovery Rate (FDR) at which a particular class (class A or class B) would stay on the discriminatory features table. This is not identical to the p-value, which is the smallest false positive rate (FPR) at which a class appears positive on the discriminatory features table. The p-value is much stricter than the q-value. An FDR of 5% (q-value <0.05) is acceptable, which is accepting 5% of erroneous single results, according to Wilcoxon test [39]. These structural features are presented below, along with sets of other correlated properties and electrostatics among classes.

### Interface polarity abundance among classes

Protein interfaces are composed of both polar and non-polar residues. Some interfaces are abundant in non-polar residues while few others are abundant in polar residues. The interface polarity abundance (P%-NP%) measure is significantly different among the interface classes with p = 7.01E-30 (**Table 1** and **Figure 2**).

### $\Delta^i$G among classes

The solvation free energy gain upon interface formation ($\Delta^i$G) is a measure of the interface stability in protein complexes [35]. The $\Delta^i$G values are significantly different among interface classes with p = 7.43E-18 (**Table 1**) as shown in **Figure 2**.

### BE among classes

The strength of binding among class A and class B complexes is estimated as a measure of BE in kcal/mol. The BE values are relatively stronger for class A complexes (average BE is -33.99 kcal/mol), as compared to class B complexes (average BE is -17.94 kcal/mol). The BE values are significantly different among interface classes with p = 2.63E-14 (**Table 1**) as shown in **Figure 2.**

### Interface charged residues among classes

The percentage of charged residues at the interface is estimated for both classes. The interface charged residues% is significantly different among interface classes with p = 6.58E-13 (**Table 1**) as shown in **Figure 2.**

### ΔASA among classes

The interface area (ΔASA) of a complex is an important structural characteristic of PPI. We observed that class A complexes demonstrate comparatively larger interfaces than class B complexes. The ΔASA is significantly different among the classes with p = 1.31E-08 (**Table 1** and **Figure 2**).

### Other correlations of interface features among classes

The stability of protein-protein binding depends on the number of hydrogen bonds and salt bridges formed between the two interacting subunits. Class A complexes show high correlation between intermolecular H-bonds and interface area (r = 0.9) as previously observed [7, 42]. However, class B complexes alone show reduced trends (r = 0.73) between intermolecular H-bonds and interface area (Additional File 3: **Figure S2)**, indicating that low quality of intermolecular hydrogen bonds is a characteristic of the large number of polar or charged residues across protein interfaces as previously observed [17]. Although salt bridges showed no distinguishing trends among classes, we observed that class B complexes are rich in salt bridges (average number of salt bridges is 6.5), as compared to class A complexes (average number of salt bridges is 5.8).

The BE values are proportional to interface area in the dataset (r = 0.96, shown in Additional File 4: **Figure S3**) as previously observed [43]. The $\Delta^{i}G$ values show relatively less correlation with interface area in class B complexes (r = -0.62) as compared to class A complexes (r = -0.92, Additional File 5: **Figure S4).** Moreover, the $\Delta^{i}G$ and BE is highly correlated among the dataset (r = 0.88) and class A (r = 0.91), however shows limited correlation among class B (r = 0.55, Additional File 6: **Figure S5**).

### Electrostatic visualization maps among protein interface classes

We have studied the surface electrostatic potential solving Poisson-Boltzmann equation using DEEPVIEW for a few examples of class A and class B complexes. This shows common surface electrostatics at work amongst the class A and amongst the class B complexes. Interestingly, the class A complexes demonstrate similar distribution of charges at the protein interfaces of both chains, suggesting electrostatic energy may not contribute to binding energy among class A complexes. However, class B complexes show opposite charge distributions at the protein interfaces, suggesting electrostatic energy plays an important role in PPIs among class B complexes as shown in **Figure 3**. Therefore, the surface electrostatic potential maps give quick visual clues for identifying or distinguishing class A and class B complexes.

## Conclusions

Structural analyses of known protein interfaces help in understanding the molecular principals of PPIs. Therefore, a comprehensive analysis of known structural interfaces of 278 complexes was carried out and their gleaned features are documented in this study. It is realized that each complex type is unique, specific and sensitive to binding. Nonetheless, there is a considerable degree of observed pattern among protein interface classes. We report two classes of interfaces, one class with less polar residues and the other class with more polar residues compared to the surfaces in bound state. The surfaces of proteins are quite polar and therefore, it is perhaps not surprising that some interfaces is polar as well and that PPI complex forms due to interactions among charged and polar residues. Thus, the need for a residue-level characterization of the interface is crucial in addition to other structural features. We document five discriminatory features (interface area, interface polarity abundance (P%-NP%), interface charged residues%, solvent free energy gain upon interface formation ($\Delta^{i}G$), and binding energy) among the interface residue-level classes. This is a first attempt towards classifying the complexes based on interface residue-level classes for the characterization of PPI features amongst these classes. These observations corroborate the need for classification of complexes in determining their combinatorial features and drawing consensus for common patterns in protein-protein recognition. These results provide molecular insights for protein-protein binding towards the development of residue-level prediction models in future studies. Additionally, mutation experiments using hot spot residue databases [44] and detailed interface residue characterization (cation-pi, electrostatic and aromatic interactions [8]) will further strengthen this study, for individual structures. Furthermore, extending this analysis for a larger dataset with a combined formulation of atomic and residue level features in future studies may improve protein-protein docking.

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

Conceived and designed the experiment: GS. Data collected, analyzed and drafted the manuscript: GS. Interpretation of results and finalizing the manuscript: GS and SR. All authors read and approved the final manuscript.

## Acknowledgments

G.S. is the recipient of an iMQRES PhD scholarship from Macquarie University. We are grateful to Dr. Edmond J. Breen for help with statistical analysis and suggestions.

## Declaration

Publication charges for this article have been funded by Macquarie University.

# References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
2. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G: **The molecular architecture of protein-protein binding sites**. *Curr Opin Struct Biol* 2007, **17**(1):67-76.
3. Keskin O, Gursoy A, Ma B, Nussinov R: **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chem Rev* 2008, **108**(4):1225-1244.
4. Sudha G, Nussinov R, Srinivasan N: **An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles**. *Prog Biophys Mol Biol* 2014.
5. Sowmya G, Ranganathan S: **Protein-protein interactions and prediction: a comprehensive overview**. *Protein Pept Lett* 2014, **21**(8):779-789.
6. Bahadur RP, Zacharias M: **The interface of protein-protein complexes: analysis of contacts and prediction of interactions**. *Cell Mol Life Sci* 2008, **65**(7-8):1059-1072.
7. Jones S, Thornton JM: **Protein-protein interactions: a review of protein dimer structures**. *Prog Biophys Mol Biol* 1995, **63**(1):31-65.
8. Gromiha MM, Yokota K, Fukui K: **Energy based approach for understanding the recognition mechanism in protein-protein complexes**. *Mol Biosyst* 2009, **5**(12):1779-1786.
9. Jones S, Thornton JM: **Principles of protein-protein interactions**. *Proc Natl Acad Sci U S A* 1996, **93**(1):13-20.
10. Gilson MK, Zhou HX: **Calculation of protein-ligand binding affinities**. *Annu Rev Biophys Biomol Struct* 2007, **36**:21-42.
11. Sowmya G, Anita S, Kangueane P: **Insights from the structural analysis of protein heterodimer interfaces**. *Bioinformation* 2011, **6**(4):137-143.
12. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites**. *Journal of molecular biology* 1999, **285**(5):2177-2198.
13. Chothia C, Janin J: **Principles of protein-protein recognition**. *Nature* 1975, **256**(5520):705-708.
14. Miller S, Lesk AM, Janin J, Chothia C: **The accessible surface area and stability of oligomeric proteins**. *Nature* 1987, **328**(6133):834-836.

15. Saha RP, Bahadur RP, Pal A, Mandal S, Chakrabarti P: **ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces**. *BMC Struct Biol* 2006, **6**:11.

16. Dong F, Zhou HX: **Electrostatic contribution to the binding stability of protein-protein complexes**. *Proteins* 2006, **65**(1):87-102.

17. Xu D, Tsai CJ, Nussinov R: **Hydrogen bonds and salt bridges across protein-protein interfaces**. *Protein Eng* 1997, **10**(9):999-1012.

18. Guharoy M, Chakrabarti P: **Conserved residue clusters at protein-protein interfaces and their use in binding site identification**. *BMC Bioinformatics* 2010, **11**:286.

19. Sheinerman FB, Norel R, Honig B: **Electrostatic aspects of protein-protein interactions**. *Curr Opin Struct Biol* 2000, **10**(2):153-159.

20. Wang T, Tomic S, Gabdoulline RR, Wade RC: **How optimal are the binding energetics of barnase and barstar?** *Biophys J* 2004, **87**(3):1618-1630.

21. Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun ZY, Wagner G, Reinherz EL: **Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors**. *Cell* 1999, **97**(6):791-803.

22. Zhang JL, Simeonowa I, Wang Y, Sebald W: **The high-affinity interaction of human IL-4 and the receptor alpha chain is constituted by two independent binding clusters**. *J Mol Biol* 2002, **315**(3):399-407.

23. Kuhlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C: **Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes**. *J Mol Biol* 2000, **301**(5):1163-1178.

24. Wong ET, Na D, Gsponer J: **On the importance of polar interactions for complexes containing intrinsically disordered proteins**. *PLoS Comput Biol* 2013, **9**(8):e1003192.

25. Sowmya G, Khan JM, Anand S, Ahn SB, Baker MS, Ranganathan S: **A site for direct integrin alphavbeta6.uPAR interaction from structural modelling and docking**. *J Struct Biol* 2014, **185**(3):327-335.

26. McCoy AJ, Chandana Epa V, Colman PM: **Electrostatic complementarity at protein/protein interfaces**. *J Mol Biol* 1997, **268**(2):570-584.

27. Kundrotas PJ, Alexov E: **Electrostatic properties of protein-protein complexes**. *Biophys J* 2006, **91**(5):1724-1736.

28. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.

29. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment**. *Proteins* 1991, **9**(1):56-68.

30. Tsodikov OV, Record MT, Jr., Sergeev YV: **Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature**. *J Comput Chem* 2002, **23**(6):600-609.

31. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility**. *J Mol Biol* 1971, **55**(3):379-400.

32. Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites**. *Proteins* 2002, **47**(3):334-343.

33. McDonald IK, Thornton JM: **Satisfying hydrogen bonding potential in proteins**. *J Mol Biol* 1994, **238**(5):777-793.

34. Gupta PS, Mondal S, Mondal B, Islam RN, Banerjee S, Bandyopadhyay AK: **SBION: A Program for Analyses of Salt-Bridges from Multiple Structure Files**. *Bioinformation* 2014, **10**(3):164-166.

35. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state**. *J Mol Biol* 2007, **372**(3):774-797.

36. Liu S, Zhang C, Zhou H, Zhou Y: **A physical reference state unifies the structure-derived potential of mean force for protein folding and binding**. *Proteins* 2004, **56**(1):93-101.

37. Zhou H, Zhou Y: **Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction**. *Protein Sci* 2002, **11**(11):2714-2726.

38. Kaplan W, Littlejohn TG: **Swiss-PDB Viewer (Deep View)**. *Brief Bioinform* 2001, **2**(2):195-197.

39. McDonald JH: **Handbook of Biological Statistics**. In., 3 edn. Baltimore, Maryland: Sparky House Publishing; 2014.

40. RStudio: **RStudio: Integrated development environment for R** In., v0.98.1091 edn. Boston; 2012.

41. Laskowski RA: **PDBsum new things**. *Nucleic Acids Res* 2009, **37**(Database issue):D355-359.

42. Gao Y, Wang R, Lai L: **Structure-based method for analyzing protein-protein interfaces**. *J Mol Model* 2004, **10**(1):44-54.

43. Chen J, Sawyer N, Regan L: **Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area**. *Protein Sci* 2013, **22**(4):510-515.

44. Ahmad S, Keskin O, Mizuguchi K, Sarai A, Nussinov R: **CCRXP: exploring clusters of conserved residues in protein structures**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W398-401.

# Table

**Table 1: PPI features distinguishing class A and class B (using Wilcoxon rank sum test)**

| PPI features | P-value | Q-value |
|---|---|---|
| Interface polarity abundance (P%-NP%) | 7.01E-30 | 1.19E-28 |
| Solvent free energy gain ($\Delta^iG$), | 7.43E-18 | 1.19E-16 |
| Binding energy | 2.63E-14 | 3.68E-13 |
| Interface charged residues% | 6.58E-13 | 8.55E-12 |
| Interface area | 1.31E-08 | 1.57E-07 |

# Figures

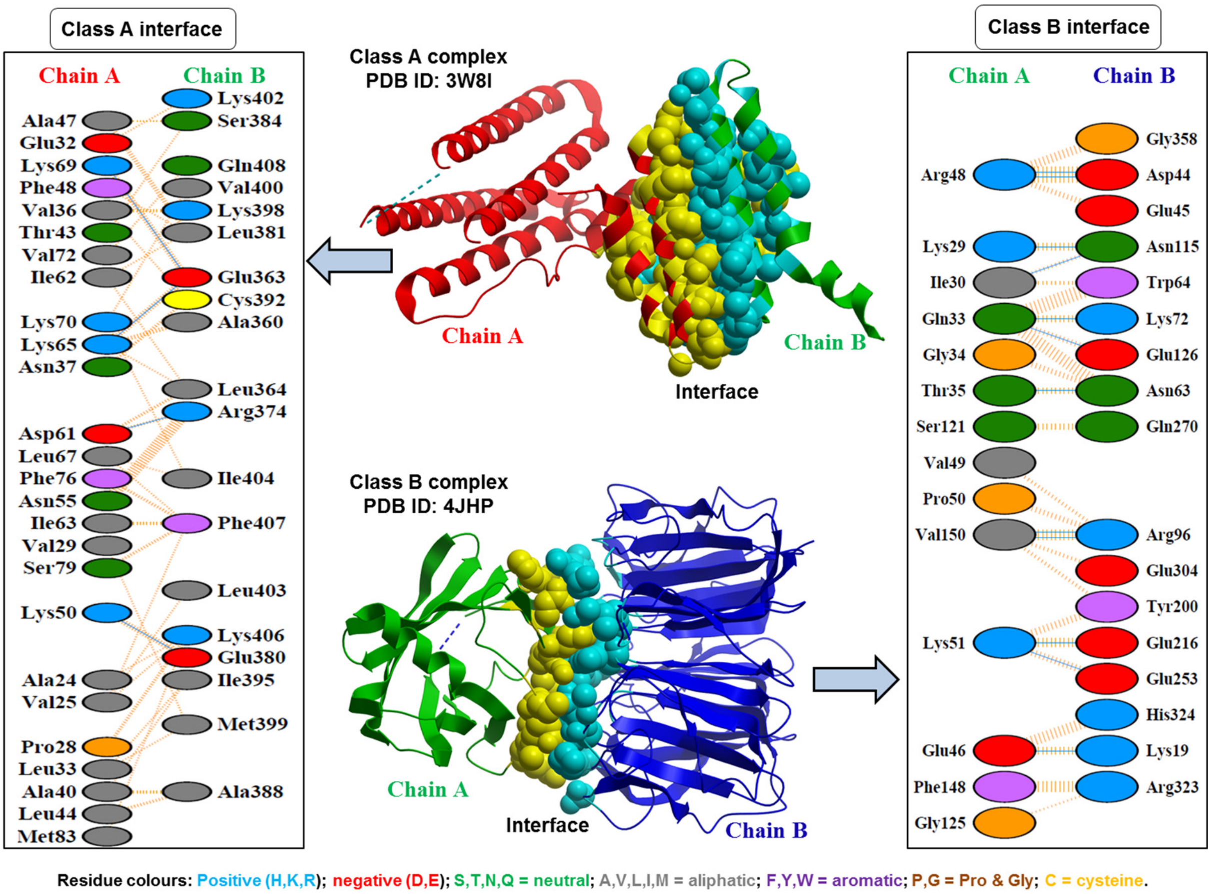


**Figure 1: Examples of PPI interfaces in class A and class B complexes.** The PDBsum [41] interaction analysis represents interaction residues on either chain with residues shown in different colours based on their properties and the coloured lines joining these residues representing the type of interaction between these residues. Class A complex shows a surface polariy of 60.28% and interface polarity of 37.84% (S>I) implying relatively less polar interactions at the interface (or relative abundance of non-polar interactions at the interface). Class B complex shows surface polariy of 50.69% and an interface polarity of 73.21% (S<I) implying relative abundance of polar interactions at the interface (or relatively less non-polar interactions at the interface).

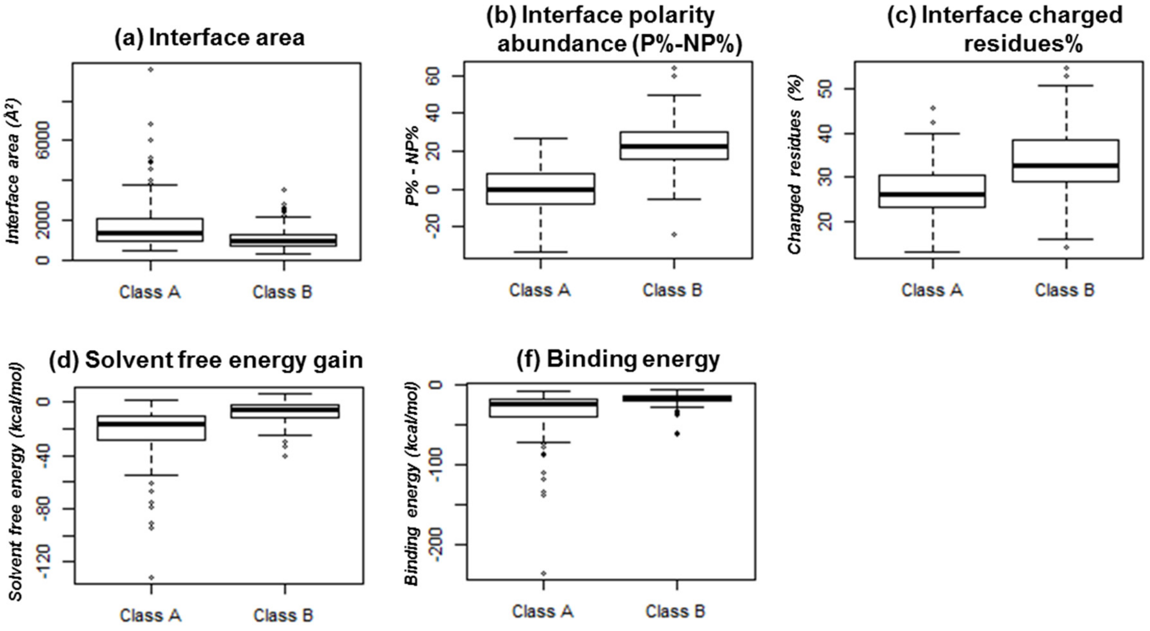


**Figure 2: Distinguishing PPI features among interface classes.** The interface area (ΔASA), interface polarity abundance (P%-NP%), interface charged residues%, solvent free energy gain ($\Delta^{i}G$), and BE are shown to distinguish among class A and class B complexes (p-values are shown in **Table 1**).

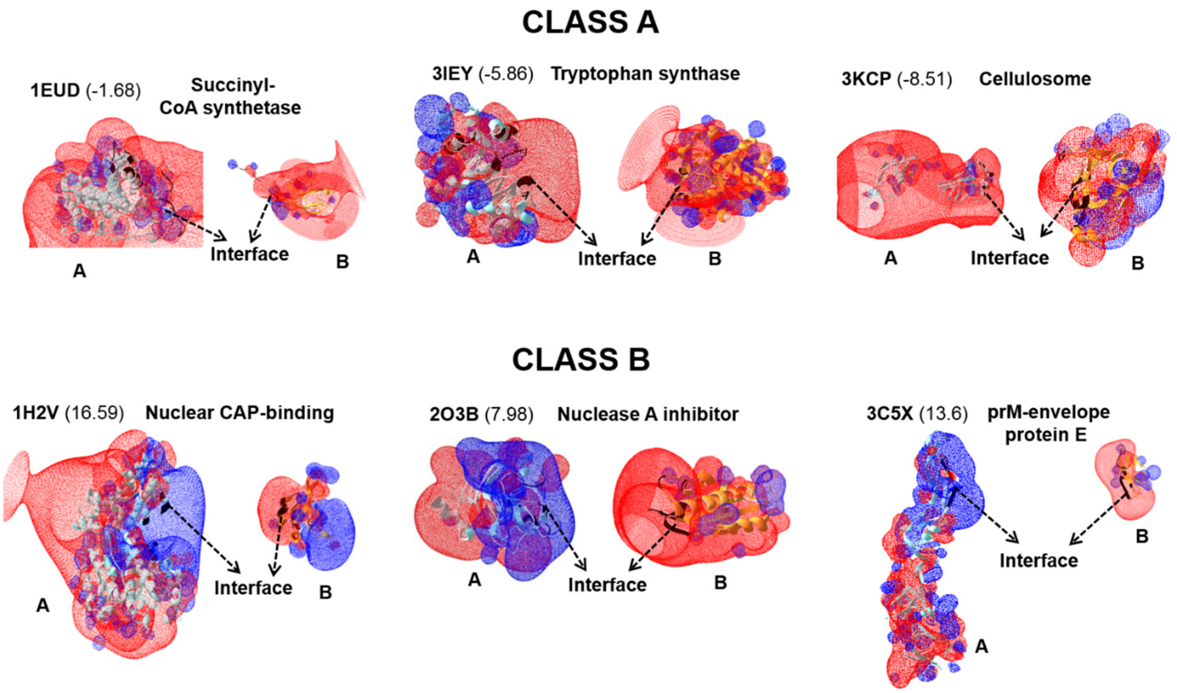

CLASS A
1EUD (-1.68) Succinyl-CoA synthetase
A
Interface
B
3IEY (-5.86) Tryptophan synthase
A
Interface
B
3KCP (-8.51) Cellulosome
A
Interface
B
CLASS B
1H2V (16.59) Nuclear CAP-binding
A
Interface
B
2O3B (7.98) Nuclease A inhibitor
A
Interface
B
3C5X (13.6) prM-envelope protein E
Interface
B
A

**Figure 3: Surface electrostatics distribution of Class A and Class B complexes using DEEPVIEW.** The heterodimer complexes are shown in cartoon representation with chain A in cyan, chain B in orange and interface residues colored in black. PDB IDs and protein names are given for each complex along with I-S values (numbers in parenthesis represent (P%-NP%), the interface polarity abundance). The electrostatic potential images of class A complexes show that the interface of chain A and of chain B have same charges (similar colors), suggesting electrostatic energy may not favor protein binding in class A complexes. The electrostatic potential images of class B complexes show that the interface of chain A and chain B have opposite charges (different colors); suggesting electrostatic energy favors protein binding in class B complexes.

# Additional Files

## Additional file 1

**Table S1: Heterodimer dataset (278) divided into interface classes based on residue level surface and interface polarity values.** The PDB code is shown along with the specific chains used in this study. S – Surface polarity; I – Interface polarity

| Class A (165) [S>I] | | | | |
|---|---|---|---|---|
| 1E44 [A, B] | 1EUD [A, B] | 1FS0 [E, G] | 1GK9 [A, B] | 1H32 [A, B] |
| 1JEQ [A, B] | 1JKG [A, B] | 1JMA [B, A] | 1LSH [A, B] | 1N1J [A, B] |
| 1NME [A, B] | 1NRJ [A, B] | 1OF5 [A, B] | 1OO0 [A, B] | 1ORY [A, B] |
| 1R0R [E, I] | 1R8O [A, B] | 1UGH [E, I] | 1US7 [A, B] | 1V74 [A, B] |
| 1VRA [A, B] | 1WPX [A, B] | 1WQJ [B, I] | 1XEW [X, Y] | 1XOU [A, B] |
| 1YKH [A, B] | 1Z0J [A, B] | 1Z5Y [D, E] | 2CG5 [A, B] | 2D74 [A, B] |
| 2DYO [A, B] | 2F4M [A, B] | 2FH5 [A, B] | 2FTX [A, B] | 2G2S [A, B] |
| 2GA9 [A, D] | 2GSK [A, B] | 2H9A [A, B] | 2OMZ [A, B] | 2OZN [A, B] |
| 2P1M [A, B] | 2PA8 [D, L] | 2PQN [A, B] | 2QSF [A, X] | 2QWO [A, B] |
| 2RAW [A, B] | 2V3B [A, B] | 2V6X [A, B] | 2VN6 [A, B] | 2VSM [A, B] |
| 2WD5 [A, B] | 2XFG [A, B] | 2ZFD [A, B] | 2ZIV [A, B] | 2ZSI [A, B] |
| 3A2F [A, B] | 3A8G [A, B] | 3AA7 [A, B] | 3ABE [C, Z] | 3AON [A, B] |
| 3AQF [A, B] | 3AU4 [A, B] | 3AYH [A, B] | 3B0C [T, W] | 3B0Z [A, B] |
| 3BS5 [A, B] | 3BTP [A, B] | 3CKI [A, B] | 3CPT [A, B] | 3CQC [A, B] |
| 3CX8 [A, B] | 3DBO [A, B] | 3DGP [A, B] | 3DPL [C, R] | 3DRA [A, B] |
| 3EGV [A, B] | 3EP6 [B, A] | 3F62 [A, B] | 3FGR [A, B] | 3FPU [A, B] |
| 3GA9 [L, S] | 3H7H [A, B] | 3HZH [A, B] | 3IEY [A, B] | 3IF8 [A, B] |
| 3JTQ [A, B] | 3K1R [A, B] | 3K8P [C, D] | 3KCP [A, B] | 3KF6 [A, B] |
| 3KXC [A, C] | 3L91 [A, B] | 3LBX [A, B] | 3LF4 [A, B] | 3LQC [A, B] |
| 3M1C [A, B] | 3M7F [A, B] | 3MCB [A, B] | 3MJ7 [A, B] | 3MKR [A, B] |
| 3ML1 [A, B] | 3MXN [A, B] | 3NW0 [B, A] | 3NYB [A, B] | 3OJA [A, B] |

|  |  |  |  |  |
|---|---|---|---|---|
| 3OSS [C, D] | 3PGE [A, B] | 3PV6 [A, B] | 3Q87 [A, B] | 3QN1 [A, B] |
| 3R07 [A, C] | 3R24 [A, B] | 3REQ [A, B] | 3RGW [L, S] | 3SHG [A, B] |
| 3T5X [A, B] | 3TBI [A, B] | 3VYR [A, B] | 3VZ9 [B, D] | 3W8I [A, B] |
| 3ZET [A, B] | 3ZVQ [A, B] | 4A5U [A, B] | 4AP2 [A, B] | 4AT7 [A, B] |
| 4AWX [A, B] | 4B8A [A, B] | 4BJJ [A, B] | 4BMP [A, B] | 4C9B [A, B] |
| 4CBU [A, G] | 4CGY [A, B] | 4CT0 [A, B] | 4CXF [A, B] | 4DBG [A, B] |
| 4DEY [B, A] | 4E4W [A, B] | 4EGC [A, B] | 4ETP [A, B] | 4EUK [A, B] |
| 4EYY [R, Q] | 4F6U [A, B] | 4F9C [A, B] | 4G1M [A, B] | 4G6T [A, B] |
| 4G94 [A, B] | 4GDX [A, B] | 4GVB [A, B] | 4H4K [A, C] | 4HNX [A, B] |
| 4HPL [A, B] | 4HST [A, B] | 4HT3 [A, B] | 4I1S [A, B] | 4JE3 [A, B] |
| 4JEH [A, B] | 4KHA [A, B] | 4KMO [A, B] | 4L2I [A, B] | 4M69 [A, B] |
| 4M6W [A, B] | 4MRT [C, A] | 4NFU [A, B] | 4NQW [A, B] | 4O8Y [A, B] |
| **Class B (113) [S<I]** |  |  |  |  |
| 1A22 [A, B] | 1ARO [P, L] | 1AY7 [A, B] | 1DJ7 [A, B] | 1GL4 [A, B] |
| 1H2V [C, Z] | 1KA9 [H, F] | 1M1E [A, B] | 1NPE [A, B] | 1SVD [A, M] |
| 1T0P [A, B] | 1T6B [X, Y] | 1USU [A, B] | 1WMH [A, B] | 1XG2 [A, B] |
| 1Z3E [A, B] | 1Z92 [A, B] | 1ZBX [A, B] | 1ZHH [A, B] | 2APO [A, B] |
| 2B42 [A, B] | 2BLF [A, B] | 2CKL [A, B] | 2FCW [A, B] | 2FHZ [A, B] |
| 2FOM [A, B] | 2HDI [A, B] | 2HRK [A, B] | 2IW5 [A, B] | 2O2V [A, B] |
| 2O3B [A, B] | 2P45 [A, B] | 2PTT [A, B] | 2QC1 [A, B] | 2QKL [A, B] |
| 2V8S [E, V] | 2VDB [A, B] | 2VLQ [A, B] | 2Z5B [A, B] | 2Z64 [A, C] |
| 3A4U [A, B] | 3ANW [A, B] | 3AWU [A, B] | 3AXJ [A, B] | 3BEG [A, B] |
| 3C5X [A, C] | 3CLS [C, D] | 3D3B [A, J] | 3DI3 [A, B] | 3DLQ [I, R] |
| 3DSS [A, B] | 3EI3 [A, B] | 3F6Q [A, B] | 3FJU [A, B] | 3FMO [A, B] |
| 3FPN [A, B] | 3FQD [A, B] | 3GB8 [A, B] | 3GC3 [A, B] | 3HHM [A, B] |
| 3KLD [A, B] | 3KYJ [A, B] | 3MCA [A, B] | 3MP7 [A, B] | 3MWD [A, B] |
| 3N1M [B, C] | 3N40 [P, F] | 3N4I [A, B] | 3NV0 [A, B] | 3NVN [A, B] |
| 3NY7 [A, B] | 3O2P [A, E] | 3O3O [A, B] | 3OG6 [A, B] | 3OJM [A, B] |

| | | | | |
|---|---|---|---|---|
| 3ONA [A, B] | 3OQ3 [A, B] | 3OUN [A, B] | 3QQ8 [A, B] | 3RNQ [B, A] |
| 3SBT [A, B] | 3THO [A, B] | 3TU3 [A, B] | 3V8X [A, B] | 3VF0 [A, B] |
| 3VRD [A, B] | 3VU9 [A, B] | 3W9C [A, B] | 3WA5 [A, B] | 3ZNZ [A, B] |
| 3ZYI [A, B] | 4BI8 [A, B] | 4BL7 [A, B] | 4C2A [A, B] | 4DRI [A, B] |
| 4DVG [A, B] | 4EMJ [A, B] | 4F48 [A, B] | 4FZV [A, B] | 4G7X [A, B] |
| 4GAF [A, B] | 4GED [A, B] | 4GQ2 [M, P] | 4HFF [A, B] | 4IU2 [A, B] |
| 4IYP [A, C] | 4J38 [A, B] | 4JHP [B, C] | 4K12 [A, B] | 4KBM [A, B] |
| 4KT1 [A, B] | 4KT3 [A, B] | 4LV5 [A, B] | | |

# Additional file 2

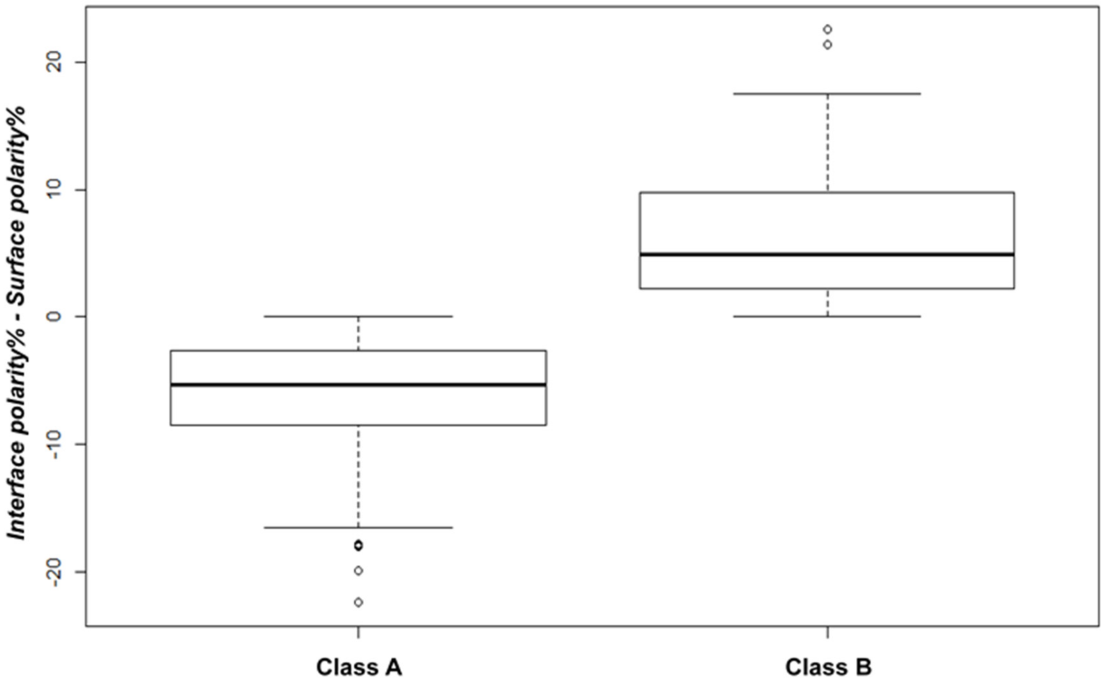


**Figure S1: Class A and Class B are significantly different.** The boxplot depicts class A and class B significantly different with a p-value of 1.66E-45 (using Wilcoxon rank sum test).

# Additional file 3

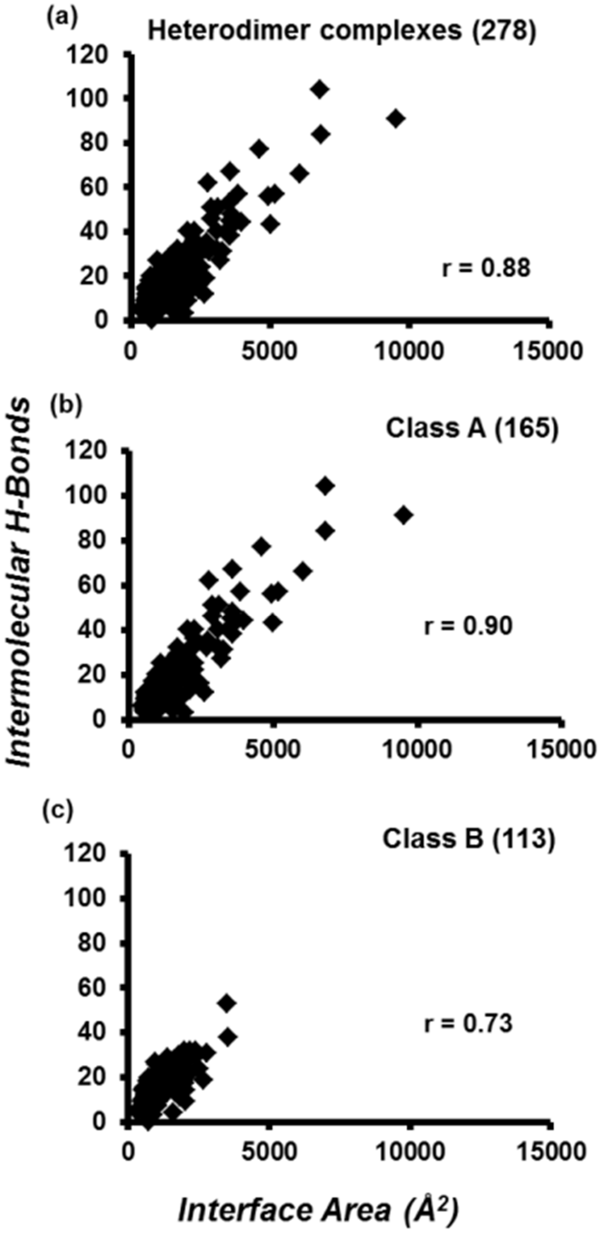


**Figure S2: Intermolecular H-bonds shows relatively low correlation with interface area in class B.** Hydrogen bonds at the protein interface are highly correlated to interface area in the dataset (r = 0.88) and class A (r = 0.9), however shows relatively lower trends (r = 0.73) in class B.

# Additional file 4

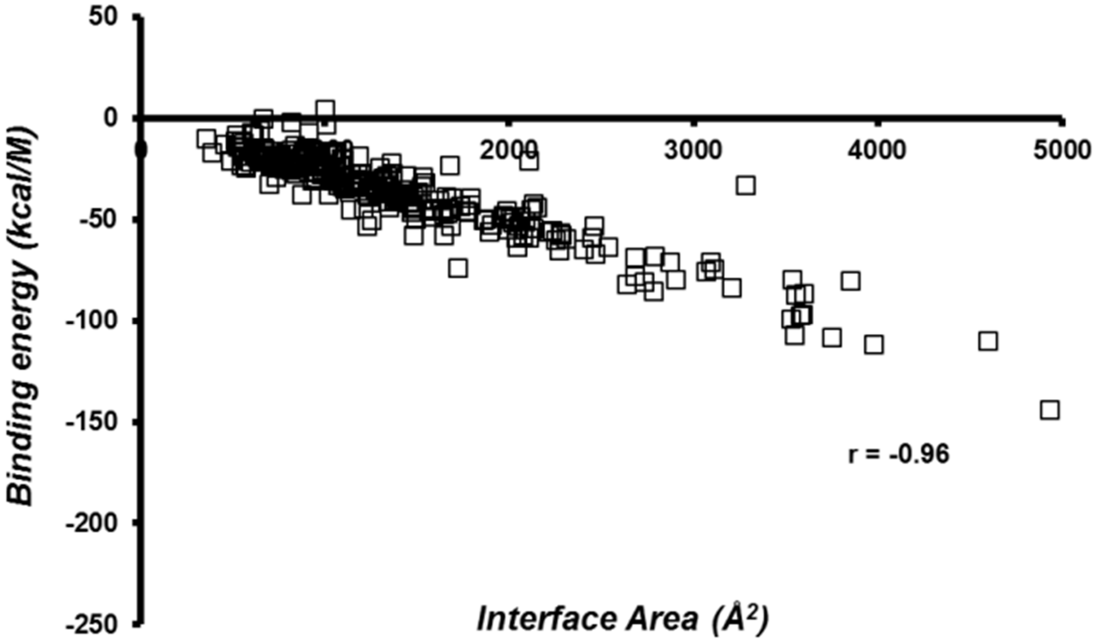


**Figure S3: Binding energy is highly correlated to interface area.** BEs at the protein interfaces are highly correlated to interface area with r = -0.96.

# Additional file 5

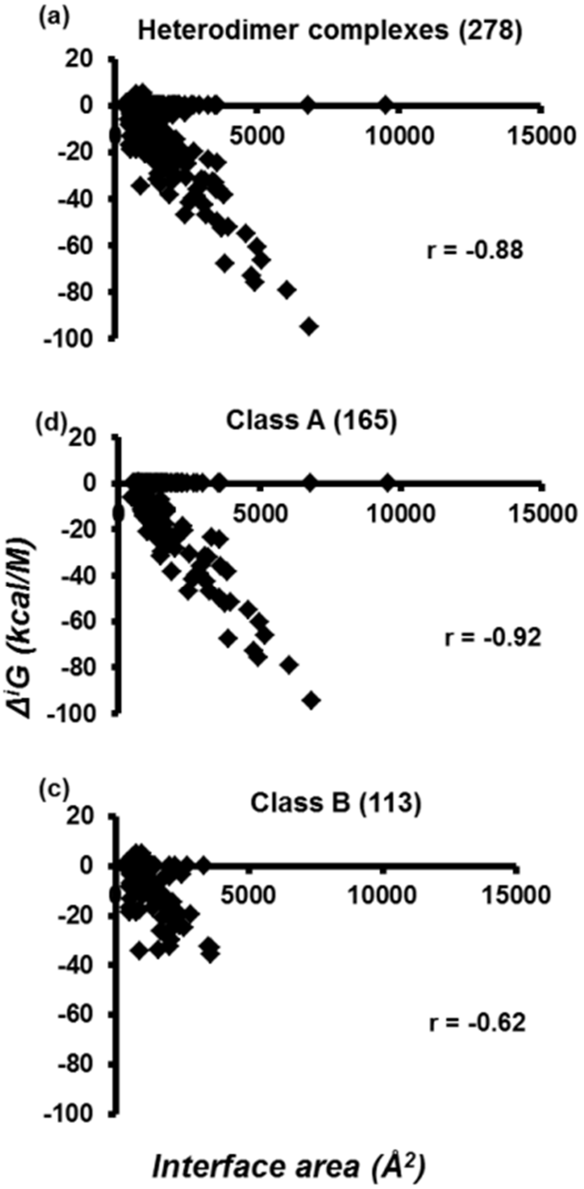


**Figure S4: Solvation free energy gain upon interface formation ($\Delta^{i}G$) shows limited correlation with interface area in class B complexes.** $\Delta^{i}G$ shows high correlation with interface area in (a) heterodimer dataset (r = -0.88), and (b) class A (r = -0.92), however shows limited correlation in (c) class B complexes (r = -0.62).

## Additional file 6

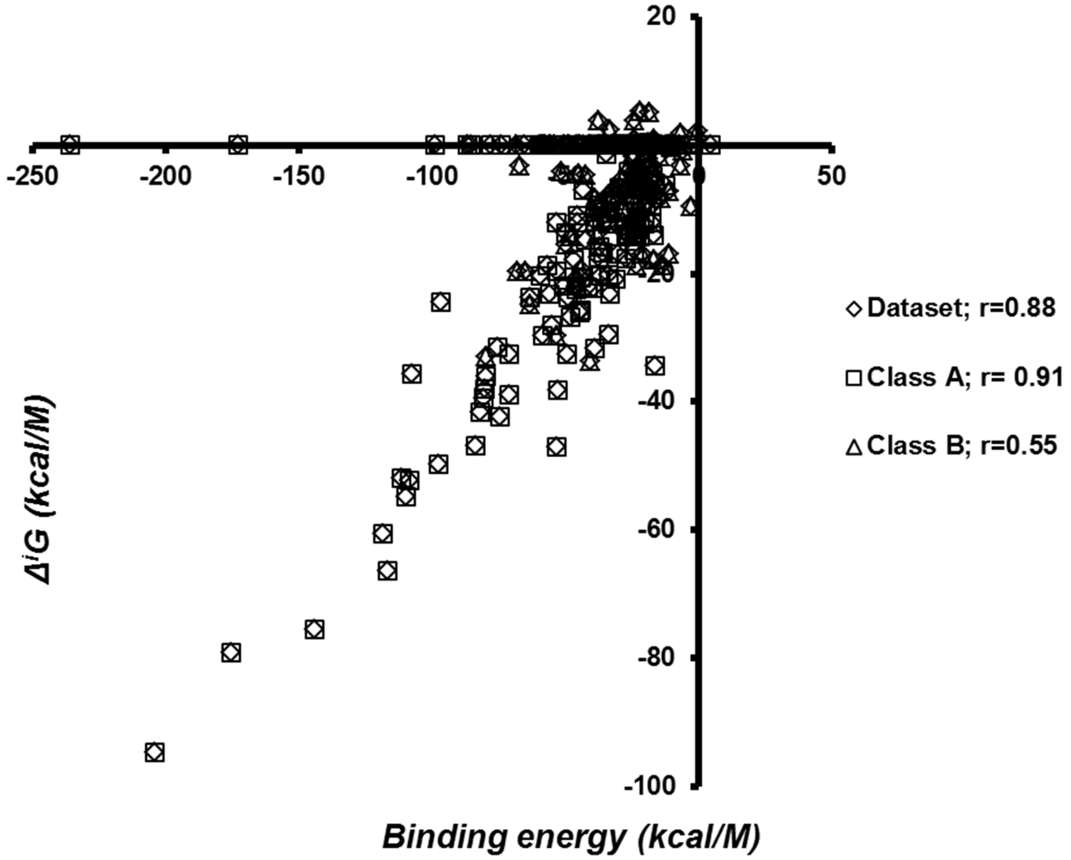


**Figure S5: BE shows limited correlated with $\Delta^i$G in class B.** Binding energies at the protein interfaces are highly correlated to solvation free energy gain upon interface formation ($\Delta^i$G) in the dataset (r = 0.88) and class A (r = 0.91), however shows limited correlation between BE and $\Delta^i$G in class B (r = 0.55).

## 3.3 Conclusions

Structural analyses of known protein-protein interfaces provide insights into understanding the major driving forces for PPI. A comprehensive structural analysis of 278 complexes is thus carried out from the PDB and documented their gleaned features in this study. It is realized that each complex type is unique, specific and sensitive to binding. Nonetheless, there is a considerable degree of observed pattern among protein interface residue-level classes. Two classes of interfaces, one class with less polar residues and the other class with more polar residues compared to their surfaces in bound state are reported.

Five key discriminatory features (interface area, interface property abundance (P%-NP%), interface charged residues%, solvent free energy gain upon interface formation ($\Delta^{i}G$), and binding energy) are identified among the interface residue-level classes. Specifically, we did not find statistically significant enhancement of any particular interface residue in our dataset. These results have application towards the development of a simple yet robust prediction model including for protein-protein binding prediction and docking studies.

Looking at the ascribed functions of protein complexes in the heterodimer dataset, all functional categories are represented in these interface classes is noted. A separate study on relating structural features to biological functions is then carried out (Chapter 4).

# Chapter 4: Protein interfaces and biological function

## 4.1 Summary

Molecular function in cellular processes is governed by PPIs. PPIs lead to diverse functionality such as catalysis, regulation, signalling, immunity and inhibition, playing a crucial role in functional genomics. However, the molecular principle of such interactions is often elusive in nature. Therefore, a comprehensive analysis of known protein complexes from the PDB is essential for the characterization of structural interface features to determine structure-function relationship.

In this study, the non-redundant dataset of 278 protein complexes described in Chapter 3, was analysed and categorized into major functional classes for distinguishing features. Several physico-chemical features such as interface size, interface area, hydrogen bonds (H-bonds), salt bridges, solvation free energy gain upon interface formation ($\Delta^{i}G$), binding energy (BE) and interface electrostatic energy ($\Delta\Delta G_{el}$) were investigated to identify discriminatory features prevailing in different functional groups. The discriminatory features among these functional groups (shown as boxplots in Figures 2 and 3 of Publication 3) along with significant correlations between these PPI features (amongst functional groups) are discussed in Publication 3.

Since Publications 2 and 3 are under consideration for different journals concurrently, some descriptive text has been repeated to make each manuscript an independent publication.

## *4.2 Publication 3*

# Linking structural features of protein complexes and biological function

Gopichandran Sowmya,[1] Edmond J. Breen,[2] and Shoba Ranganathan[1]*

[1]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia
[2]Australian Proteome Analysis Facility (APAF), Macquarie University, Sydney, New South Wales 2109, Australia

Received 1 April 2015; Accepted 24 June 2015
DOI: 10.1002/pro.2736
Published online 30 June 2015 proteinscience.org

**Abstract: Protein–protein interaction (PPI) establishes the central basis for complex cellular networks in a biological cell. Association of proteins with other proteins occurs at varying affinities, yet with a high degree of specificity. PPIs lead to diverse functionality such as catalysis, regulation, signaling, immunity, and inhibition, playing a crucial role in functional genomics. The molecular principle of such interactions is often elusive in nature. Therefore, a comprehensive analysis of known protein complexes from the Protein Data Bank (PDB) is essential for the characterization of structural interface features to determine structure–function relationship. Thus, we analyzed a non-redundant dataset of 278 heterodimer protein complexes, categorized into major functional classes, for distinguishing features. Interestingly, our analysis has identified five key features (interface area, interface polar residue abundance, hydrogen bonds, solvation free energy gain from interface formation, and binding energy) that are discriminatory among the functional classes using Kruskal-Wallis rank sum test. Significant correlations between these PPI interface features amongst functional categories are also documented. Salt bridges correlate with interface area in regulator-inhibitors ($r = 0.75$). These representative features have implications for the prediction of potential function of novel protein complexes. The results provide molecular insights for better understanding of PPIs and their relation to biological functions.**

**Keywords: protein–protein interaction; heterodimers; surface; interface; protein structure; protein function**

## Introduction

Protein–protein interaction (PPI) is critical for molecular functions in living systems. PPIs are associated with catalysis, regulation, signaling, immunity, and inhibition, thereby playing a critical role in functional genomics.[1] Extensive studies on the existing PPI complexes are the key to understanding cellular machinery as reviewed elsewhere.[1–5] With modern experimental techniques such as two hybrid systems, protein fragment complementation, tandem affinity purification methods and protein arrays, several interacting protein pairs have been detected in large-scale studies,[6] although their biological role may not be well characterized or known. Comprehensive homology modeling techniques of known interacting proteins combined with docking studies and PPI data helps in understanding structural assembly for functional preferences as shown for integrin αvβ6 heterodimer (two different protein subunits) complex.[7] Thus, the analysis of PPIs is essential in predicting important biological functions.[8,9] It is well known that structure-based characterization of multimeric proteins is the key to ascribing biological functional annotation.[10]

The availability of protein–protein complexes at the Protein Data Bank (PDB)[11] has enabled sequence and structural investigations on the existing complexes

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Shoba Ranganathan, Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia.
E-mail: shoba.ranganathan@mq.edu.au

to decipher their recognition mechanisms and binding principles for decades.[3,12–15] The classical work by Chothia and Janin[16] with three protein complexes defined 'hydrophobicity' (nonpolar interactions) as the major stabilizing factor in PPI, which has been affirmed with larger datasets of protein complexes.[12,14,17–19] Conversely, shape complementarity, polar interactions, hydrogen bonding, and salt bridges are believed to primarily contribute to binding specificity and free energy of binding.[20–23] Furthermore, a number of physicochemical factors known to govern protein–protein association include interface size, planarity, sphericity, complementarity, types of amino acid chemical groups, hydrophobicity, electrostatic interactions, H-bonds, distribution of binding energy, sequence conservation, and conserved residue clusters.[12,16,21,24–27] However, extensive surveys carried out thus far by various groups' typically average structural features over diverse datasets of protein–protein complexes, obscuring information on individual proteins' structural integrity. In an earlier study, we developed a homology model for integrin αvβ6 heterodimer[7] where we found that the subunit interface in our model as well as the template X-ray structures clearly showed an increase of polar residues compared with the surface or the complex. Therefore, it is essential to revisit interface analysis to understand PPI binding principles, using a large nonredundant structural dataset.

The classification of protein–protein complexes based on their composition, affinity, interface stability, and lifetime association into different groups has gained momentum in the past decade,[28–30] although the boundaries between these classes is often indefinite, based on physiological conditions.[31] Alternatively, the classification of complexes into major functional groups can be valuable in relating structural data to biological functions for better understanding of PPIs.[32] Moreover, the classification of protein–protein complexes based on functions and their usefulness in improving prediction accuracy has been observed recently.[33] Therefore, for analyzing PPI primarily based on functions, a nonredundant set of complexes, with structural information, is essential.

Nonidentical protein subunits (or chains) noncovalently interact to form 'heteromers' with diverse functionality. Heterodimeric interactions are commonly found in enzyme-inhibitors, enzyme complexes, antibody–antigen, signal proteins and cell cycle proteins, and also include transient complexes. Dimeric interactions are amongst the strongest and most extensive in nature.[12,14] In this study, we have created a nonredundant dataset of 278 heteromeric protein structure dimers from the PDB, for characterization of their interface structural features to determine whether biochemical function is related to interface features. Several physico–chemical features such as interface size, interface area, hydrogen bonds (H-bonds), salt bridges, solvation free energy gain ($\Delta^{i}G$), binding energy (BE), and interface electrostatic energy ($\Delta\Delta G_{el}$) were investigated to study discriminatory features prevailing in different functional groups. Our analysis has identified key features [interface area, interface polarity abundance (P% − NP%), H-bonds, $\Delta^{i}G$, and BE] that are significantly different between the functional groups. This result has implications for function prediction for orphan proteins, where interacting partners are known and heteromeric complexes can be structurally modeled with high confidence.

## Results and Discussion

Our nonredundant dataset (described in Materials and Methods: Heterodimer dataset) comprised 278 complexes (Table I). Our previous dataset is 98.7% of the current dataset with only five new entries in the current dataset (underlined in Table I). The dataset was then grouped into important functional groups and protein types to better understand PPIs among these divisions of complexes. The dataset comprises enzymes (40), regulators (144), enzyme-inhibitors (25), regulator-inhibitors (27) and immune (18) and biological assembly (24) complexes, of which 247 are globular and 31 membranous. The distribution of complexes under different functional classes is unbiased. Enzyme complexes are formed when two enzyme subunits interact to achieve a 'catalytic' function. For example, PDB ID: 1EUD succinyl coenzyme A (succinyl-CoA) synthetase is formed by the interaction between succinyl CoA synthetase α chain (311 AA) and succinyl-CoA synthetase β chain (396 AA). The succinyl-CoA synthetase (SCS) protein catalyses a reversible conversion of succinyl-CoA and succinate, coupled with phosphorylation/dephosphorylation. Enzyme complexes considered in this study do not include enzyme-inhibitor complexes. Enzyme-inhibitor complexes are formed when an enzyme and an inhibitor protein interact to achieve an 'inhibitory' function. For example, PDB ID: 1ARO is T7 RNA polymerase (883 AA) complexed to the transcriptional inhibitor T7 lysozyme (151 AA). The lysozyme binds at a site distant from the polymerase active site, suggesting an indirect mechanism of inhibition.

Regulatory complexes are formed between two protein subunits to achieve a 'regulatory' or 'cellular' function. For example, PDB ID: 1JEQ is a Ku heterodimer, formed by Ku70 (609 AA) and Ku80 (565 AA) subunits, contributing to genomic integrity by binding to DNA double-strand breaks and enabling repair by nonhomologous end-joining pathway. The regulatory complexes do not include regulator-inhibitor complexes, which are formed between a regulatory protein subunit and an inhibitor protein to achieve an 'inhibitory' function. For example, PDB ID: 1A22 is a complex between the G120R mutant (191 AA) of the human growth hormone and its cognate receptor, the human growth hormone receptor (238 AA). The G120R mutant of human

**Table I.** *A Dataset of Heterodimer Protein Complexes (278) Divided into Literature Driven Functional Groups with Major Protein Types. New entries added to the earlier dataset [42] are underlined.*

| Functional groups | Globular (247) | | | | | | | Membrane (31) |
|---|---|---|---|---|---|---|---|---|
| Enzymes (40) | 1EUD | 2APO | 2O2V | 3EP6 | 3O3O | 3REQ | 4HST | 3RGW |
| | 1GK9 | 2BLF | 2QKL | 3FGR | 4EMJ | 3ZVQ | 4MRT | |
| | 1NME | 2CG5 | 3A8G | 3GA9 | 3L91 | 4BMP | 4NFU | |
| | 1VRA | 2GA9 | 3AON | 3JTQ | 3M7F | 4DBG | | |
| | 1KA9 | 2XFG | 3AYH | 3DSS | 3ML1 | 4GED | | |
| | 1SVD | 2ZIV | 3DRA | 3MWD | 3R07 | 4HNX | | |
| Regulators (144) | 1JEQ | 2G2S | 3B0C | 3LF4 | 4AT7 | 4G6T | 4IYP | 1FS0 |
| | 1JKG | 2H9A | 3B0Z | 3MCB | 4AWX | 3VRD | 4JHP | 1H32 |
| | 1LSH | 2HRK | 3CPT | 3MKR | 4BJJ | 3W9C | 4KBM | 1NRJ |
| | 1N1J | 2IW5 | 3CX8 | 3MXN | 3O2P | 3ZNZ | 4KT1 | 1Z0J |
| | 1OF5 | 2VDB | 3DGP | 3NW0 | 3OG6 | 3ZYI | | 1Z5Y |
| | 1OO0 | 2P1M | 3DPL | 3GC3 | 3OJM | 4BL7 | | 2GSK |
| | 1ORY | 2QSF | 3EGV | 3HHM | 3SBT | 4C2A | | 2PQN |
| | 1US7 | 2QWO | 3CLS | 3KLD | 3THO | 4DVG | | 2V6X |
| | 1WQJ | 2V3B | 3D3B | 3KYJ | 3TU3 | 4GDX | | 3AQF |
| | 1YKH | 2VN6 | 3EI3 | 3MCA | 3VF0 | 4JE3 | | 3BS5 |
| | 2D74 | 2WD5 | 3F6Q | 3N1M | 4C9B | 4JEH | | 4CXF |
| | 1DJ7 | 2ZFD | 3FMO | 3NV0 | 4CBU | 4KHA | | 4G1M |
| | 1H2V | 2ZSI | 3FQD | 3OSS | 4CGY | 4KMO | | 1ZHH |
| | 1USU | 3ANW | 3GB8 | 3PGE | 4CT0 | 4L2I | | 2FCW |
| | 1WMH | 3AWU | 3F62 | 3Q87 | 4DEY | 4M69 | | 2V8S |
| | 1Z3E | 3AXJ | 3H7H | 3T5X | 4E4W | 4M6W | | 3MP7 |
| | 1ZBX | 3BEG | 3HZH | 3TBI | 4EGC | 4NQW | | 3NY7 |
| | 2CKL | 3A2F | 3IF8 | 3VZ9 | 4ETP | 4O8Y | | 3OUN |
| | 2DYO | 3ABE | 3K8P | 3ZET | 4EUK | 4FZV | | 3V8X |
| | 2FH5 | 3AU4 | 3KXC | 4A5U | 4EYY | 4GQ2 | | 4G7X |
| Enzyme-inhibitors (25) | 1ARO | 1AY7 | 2OZN | 3QN1 | 3N4I | 4DRI | 4LV5 | |
| | 1R0R | 1WPX | 3CKI | 2O3B | 3SHG | 4F9C | | |
| | 1R8O | 1XG2 | 3DBO | 2VLQ | 4F6U | 4HT3 | | |
| | 1UGH | 2B42 | 3IEY | 3R24 | 3FJU | 4I1S | | |
| Regulator-inhibitors (27) | 1A22 | 2FOM | 2QC1 | 3NVN | 3OQ3 | 4BI8 | 4GVB | 1E44 |
| | 1M1E | 2HDI | 3N40 | 3ONA | 3QQ8 | 4GAF | | 1T6B |
| | 1JMA | 2F4M | 2RAW | 3AA7 | 3NYB | 4B8A | | |
| | 1XOU | 2OMZ | 2VSM | 3BTP | 3OJA | 4G94 | | |
| Immune complexes (18) | 1Z92 | 2PTT | 3DLQ | 3PV6 | 4HFF | 4KT3 | | 1T0P |
| | 2FHZ | 2Z64 | 3FPU | 3RNQ | 4J38 | | | 1V74 |
| | 2P45 | 3DI3 | 3MJ7 | 3WA5 | 4K12 | | | |
| Biological assembly (24) | 1XEW | 2Z5B | 3K1R | 3LQC | 3W8I | 4H4K | | 1GL4 |
| | 2FTX | 3CQC | 3KCP | 3VYR | 4AP2 | 4HPL | | 1NPE |
| | 2PA8 | 3FPN | 3KF6 | 3VU9 | 4F48 | 4IU2 | | 3A4U |
| | | | | | | | | 3C5X |
| | | | | | | | | 3LBX |
| | | | | | | | | 3M1C |

growth hormone (hGH) is an antagonist binding to the growth hormone receptor. An example for each of these functional groups is shown in Figure 1.

### PPI features among functional groups

Of all the physicochemical features described in the Materials and Methods section, five were significant in discriminating between the functional categories. The results from these features, measured in terms of six parameters (Fig. 2), using a Kruskal-Wallis rank sum test (*P*-values in Table II) are presented below, along with sets of correlated properties.

### Interface area among complexes

The standard value reported by Lo Conte and colleagues for a protein interface is 1600 Å$^2$ (±400).[26] Bahadur and colleagues showed that the range of interface area (ΔASA) extends from 500 to 7000 Å$^2$, with a mean of 1970 Å$^2$,[34] while Caffrey *et al.* showed that interface area ranges from 415 to 2361 Å$^2$ for heterodimer complexes.[35] Our analysis (described in Materials and Methods: Interface analysis) shows that the interface area per subunit ($B/2$) ranges from ~350 to 9500 Å$^2$ (±1100) and differs for each functional group as shown in Figure 2. The interface area is significantly different among functional groups with $P = 2.25\text{E}{-}05$ (Table II).

### Interface polarity abundance

The difference in percentages between interface polar residues and nonpolar residues (P% − NP%) gives the measure of the abundance of polar or nonpolar residues at the interface.[32] The interface polarity abundance (P% − NP%) measure (described in

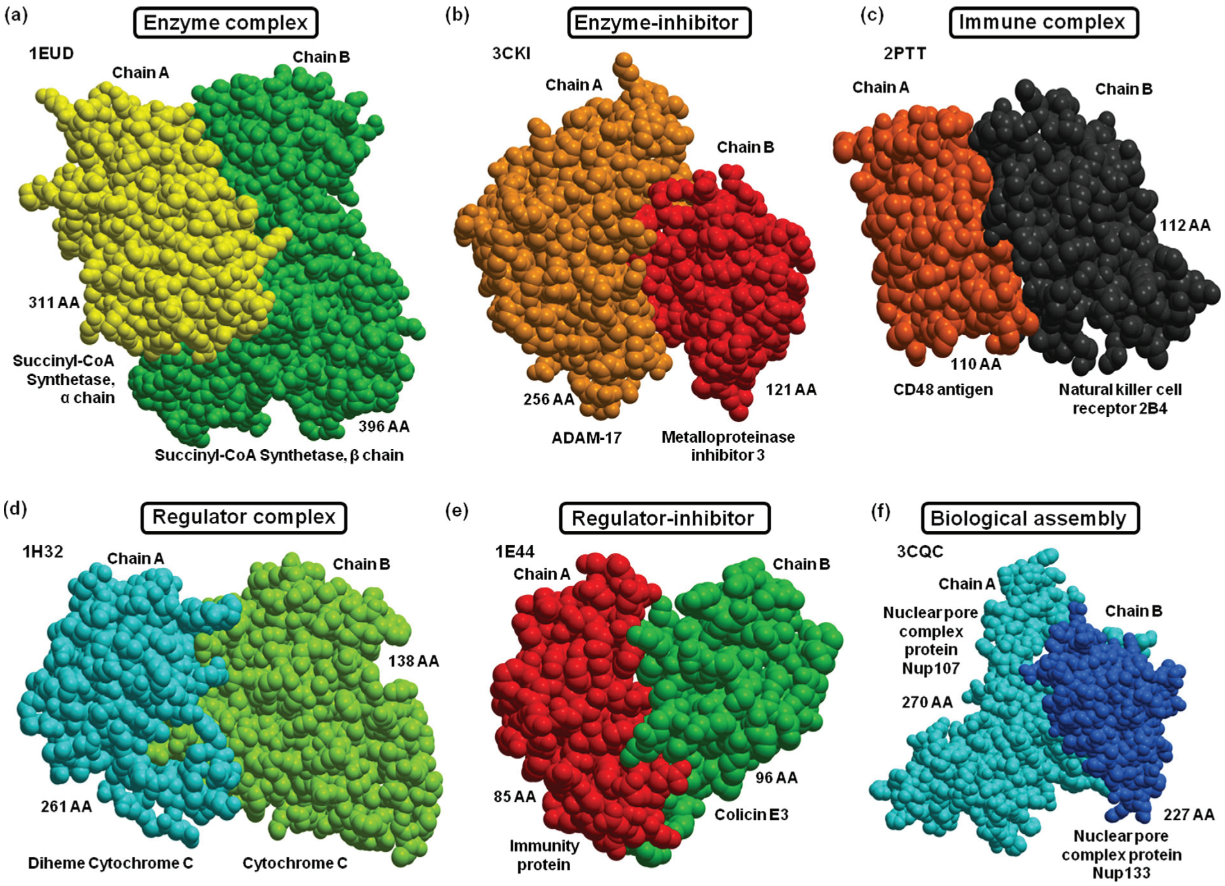


**Figure 1.** Examples of protein–protein complex structures from each functional group. (a) An enzyme complex (PDB: 1EUD) formed between succinyl-CoA synthetase α chain and succinyl-CoA synthetase β chain is shown in yellow and green colors, respectively. (b) TACE-N-TIMP-3 enzyme-inhibitor complex (PDB: 3CKI) with ADAM (a disintegrin and metalloproteinase) inhibition by TIMP-3 (tissue inhibitor of metalloproteinases 3) is shown in light brown and red, respectively. (c) An immune complex (PDB: 2PTT) between NK cell receptor 2B4 (CD244) bound to CD48 is shown in black and dark brown, respectively. (d) A regulatory SoxAX protein (involved in enzymatic oxidation of thiosulfate; PDB:1H32) of the diheme cytochrome C (in cyan colored) and cytochrome C (in light green) is shown. (e) A regulator–inhibitor (PDB:1E44) of the cytotoxic domain of colicin E3 in complex with its immunity protein is shown in green and red, respectively. (f) A biological assembly of the nuclear pore complex (NPC) between Nup107 and Nup133 (PDB: 3CQC) is shown in cyan and dark blue, respectively.

Materials and Methods: Interface analysis) is significantly different among the different functional categories with $P = 4.25\text{E}{-}05$ (Table II and Fig. 2).

### Hydrogen bonds among complexes

The stability of protein–protein binding depends on the number of hydrogen bonds (H-bonds) and salt bridges formed between the two interacting subunits. On an average, 10.1 H-bonds are formed at a protein–protein interface, with one H-bond per 170 Å$^{2}$ interface area with an $r$ value of 0.84 observed between H-bonds and interface area.[26] The $r$ value between H-bonds and interface area calculated using different dataset size and nature of data varies from 0.75 to 0.89,[12,14,21,26,36,37] with an average of 0.24 H-bonds per interface residue in heterodimers. High H-bond density per interface residue (0.64) with dominant charged and hydrophilic/polar residues at the heterodimer protein interfaces is also demonstrated.[37] Our statistical analysis (described in Materials and Methods: Intermolecular hydrogen bonds calculation) shows that the total number of intermolecular H-bonds for each functional group is significantly different with $P = 7.54\text{E}{-}04$ (Table II and Fig. 2).

### Solvent free energy gain upon interface formation

The solvation free energy gain upon interface formation ($\Delta^{i}G$)) [calculated as described in Materials and Methods: calculation of solvation free energy gain upon interface formation ($\Delta^{i}G$)] is significantly different among functional groups with $P = 7.08\text{E}{-}06$ 04 (Table II) as shown in Figure 2.

### Binding energy at the interface

To study the strengths of binding among functional groups, we estimated the binding free energy (BE; also called binding affinity calculated as described in Materials and Methods: Binding energy calculation)

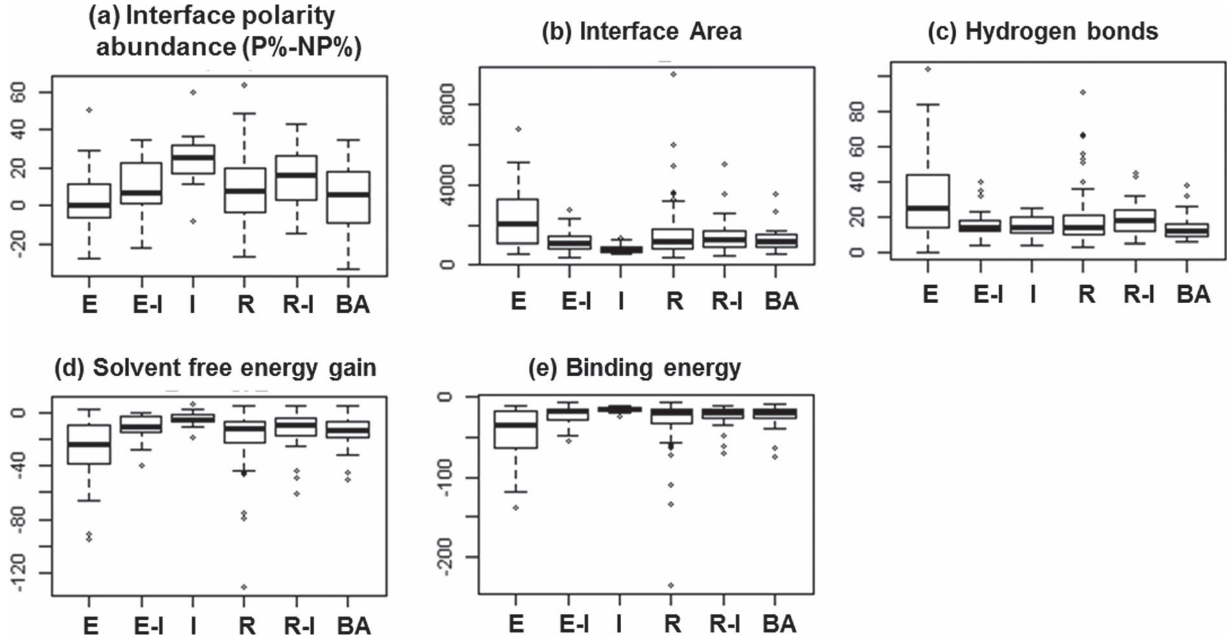


**Figure 2.** Discriminatory structural PPI features among functional groups. The interface polarity abundance (P% − NP%), interface area, hydrogen bonds, solvent free energy gain, and binding energies are significantly different among functional groups obtained using Kruskal-Wallis test (*P*-values shown in **Table II**). The functional classes include E, enzymes; E-I, enzyme-inhibitors; I, immune; R, regulators; R-I, regulator-inhibitors; BA, biological assembly.

of the dataset. The BEs are significantly different among functional groups with $P = 3.11\text{E}{-}05$ (Table II) as shown in Figure 2.

### *Correlations amongst interface physiochemical features*

Interface polarity abundance (P% − NP%) shows limited correlation with charged residues at the interface ($r = 0.61$) for the heterodimer dataset. However, this is not unexpected as charged residues are included in the polar residue set. Also, the protein complexes in our dataset show high correlation between intermolecular H-bonds and interface area ($r = 0.90$) as previously observed.[12,14,21,26,36,37] Moreover, H-bonds are correlated to BE ($r = -0.70$). Salt bridges (described in Materials and Methods: Intermolecular salt bridges calculation) across the interface do not show any significant correlation to all other features, which is in accord with another study.[21] Although this parameter was not statistically significant between the different functional groups, it is correlated to interface area in regulator-inhibitor complexes ($r = 0.75$; see Supporting Information Fig. S1). This shows that intermolecular salt bridges are an important structural feature in some functional complexes.

The solvation free energy gain upon interface formation ($\Delta^{i}G$) shows high correlation with interface area in the heterodimer dataset ($r = -0.88$), suggesting $\Delta^{i}G$ is an important feature in characterizing PPIs. Interestingly, biological assembly and immune complexes showed the least correlation of $\Delta^{i}G$ with interface area ($r = -0.67$) as compared with other functional groups as shown in Supporting Information Figure S2. The $\Delta^{i}G$ and BE are also highly correlated among the dataset ($r = 0.88$). Our analysis shows a high correlation between BE and interface area in the dataset ($r = 0.96$) as previously observed.[38]

The electrostatic component of binding free energy ($\Delta\Delta G_{\text{el}}$) was studied to quantify the electrostatic free energy favoring protein–protein interaction among functional groups of complexes (described in Materials

**Table II.** *Discriminatory PPI Features Among Functional Groups (Using Kruskal-Wallis Rank Sum Test)*

| Functional groups | *P*-value | *Q*-value |
|---|---|---|
| Solvent free energy gain ($\Delta^{i}G$) | 7.08E−06 | 0.000128 |
| Interface area | 2.25E−05 | 0.000382 |
| Binding energy | 3.11E−05 | 0.000497 |
| Interface polarity abundance (P%−NP%) | 4.25E−05 | 0.000638 |
| Hydrogen bonds | 7.54E−04 | 0.009964 |
| Solvent free energy gain ($\Delta^{i}G$) *P*-value | 7.12E−04 | 0.009964 |

*Abbreviations*: $\Delta^{i}G$, solvent free energy gain; ΔASA, interface area; $\Delta\Delta G_{\text{el}}$, interface electrostatic free energy calculation; BE, binding energy; H-bonds, hydrogen bonds; ICM, internal coordinate mechanics; NP, nonpolar residues; PDB, Protein Data Bank; P, polar residues; PPI, protein–protein interaction.

and Methods: Interface electrostatic free energy calculation ($\Delta\Delta G_{el}$)). Interface electrostatic energy component of BE (Supporting Information Fig. S3) shows distribution of charges in the dataset with electrostatic energy contributing to destabilizing PPIs in a few complexes while stabilizing PPIs in the others, suggesting that quantification of accurate interface electrostatic component contributing to BE is often a nontrivial task. The frequency distribution also shows similar trends for interface electrostatic energy (Supporting Information Fig. S4). Although $\Delta\Delta G_{el}$ values show limited correlation with interface area ($r = -0.47$), the enzymes complexes show correlation between $\Delta\Delta G_{el}$ and interface area ($r = -0.61$), as opposed to other groups ($r < 0.5$).

These discriminatory PPI features hold significantly different among functional groups in globular proteins with $P < 0.05$ as shown in supplementary Supporting Information Table S1. Since there is insufficient information regarding all the functional groups in membrane proteins, the discriminatory features among functional groups in membrane proteins is not clear at this point. These observations corroborate the need for classification of complexes in determining their combinatorial features and drawing consensus for common patterns in protein–protein recognition. Incorporation of these combinatorial features among protein functional groups is necessary to develop models for residue-level protein–protein binding prediction and analysis, and also in utilizing PPI information for the prediction of potential protein functions in future studies.

## Materials and Methods

### *Heterodimer dataset*

We created a nonredundant protein heterodimer (comprising two different protein subunits) dataset, the 3D structures of which were determined by X-ray crystallography, from the PDB, using the RCSB's advanced search interface. The search criteria were: (i) resolution <= 3Å and (ii) protein size > 50 residues, as described in earlier studies (iii) limited to "experimental data" to obtain high resolution true structures and avoid short peptides, synthetic, or artificial complexes and (iv) the number of chains, entities, and oligomeric state is set at two to obtain dimers with two unique or different chains, (v) exclude DNA or RNA or a hybrid of such molecules with proteins or otherwise and (vi) sequence identity cut-off is set to 30% and (vii) the select parameter was set so that entries with mutations were not included in the dataset. The redundancy among heterodimer complexes was further removed using the USEARCH program[39] at sequence identity cut-off of 20%, as this threshold eliminates remote homology of 25%[40] seen in structures as well. The criteria set out are comparable with those of Janin and coworkers,[41] who have used a resolution cutoff for X-ray structures (3.25Å) and chain length (minimum of 30 residues) for defining a benchmark dataset for complexes with experimental binding energies, toward developing a method for binding energy prediction.

### *Grouping based on function and protein type*

The assembly of proteins into functional complexes is essential in biology. Therefore, the characterization of these functional complexes is an essential step in deciphering their binding principles. Hence, we have grouped the dataset into major functional groups such as enzymes (E), enzyme-inhibitors (EI), regulators (R), regulator-inhibitors (RI), immune (I), and biological assembly (BA) complexes, as described in the PDB header. In a few cases, where the complexes had more than one function, the functional group assigned is based on their primary role obtained from literature. The dataset comprised globular (247) and membranous (31) complexes.

## Structural Analysis

### *Interface analysis*

A protein–protein interface is identified by calculating the change in solvent accessible surface area ($\Delta$ASA) upon binding.[42] Surface Racer 5.0 program[43] with a probe radius of 1.4Å and Lee and Richards implementation[42] was used to calculate ASA. $\Delta$ASA (interface area) of a heterodimer complex was calculated as shown in Eq. (1).

$$\Delta\text{ASA(complex)} = [\text{ASA(subunit A)} + \text{ASA(subunit B)} - \text{ASA(complex AB)}/2 \tag{1}$$

The identified interface residues were further filtered based on the criteria that their $\Delta\text{ASA} > 0.1\text{Å}^2$.[44] The amount of polar, nonpolar, and charged residues at the interface was then estimated for the dataset. Polar residues considered in the analysis are R, N, D, E, Q, H, K, S, T, and Y.[45] We have computed the relative abundance of polar or nonpolar residues ["interface polarity abundance (P% − NP%)"] compared with nonpolar residues at the interface as the difference between the percentage of interface polar residues and the percentage of interface nonpolar residues.

### *Intermolecular hydrogen bonds calculation*

The hydrogen atoms covalently bound between two electronegative atoms and contributing to electrostatics was calculated using HBPLUS program.[46] The output file of HBPLUS contains information on all donor and acceptor atoms, angles and distances within the distance of 4 Å and then filtered for intermolecular hydrogen bonds, where the hydrogen bond donor and acceptor atoms are from two different subunits.

### Intermolecular salt bridges calculation

The salt bridges formed between two oppositely charged side-chain atoms, (i.e., basic and acidic amino acids) within a distance of 4 Å and contributing to the stability and electrostatics of the protein complex was calculated using the SBION program,[47] and intermolecular salt bridges were then extracted such that the oppositely charged atoms are from two different subunits.

### Calculation of solvation free energy gain upon interface formation ($\Delta^iG$)

The $\Delta^iG$ of protein complex arises from the change in solvation energy as well as contact-dependent and electrostatic interactions of the subunits, quantifying the solvation free energy gained upon interface formation. The difference in total solvation energies between the isolated and complexed structure gives the solvation free energy gain at the interface (in kcal/mol), The PDBePISA webserver[48] was used to calculate $\Delta^iG$ for the heterodimer dataset.

### Binding energy calculation (BE)

The interaction between two protein subunits can be characterized in terms of binding free energy. The BE term (comprising electrostatic, van der Waals, hydrophobic, and entropic terms) gives an indication of strong and weak intermolecular forces, with the most negative value considered the strongest. The binding free energy of 278 protein complexes was calculated using the DCOMPLEX program,[49] which uses DFIRE-based potentials.[50] DCOMPLEX is the most widely used program for calculating binding energies of protein complexes as it has been benchmarked to reproduce experimental binding free energy values.

### Interface electrostatic free energy ($\Delta\Delta G_{el}$) calculation

Interface electrostatic free energy component of the binding free energy ($\Delta\Delta G_{\mathrm{el}}$) was calculated as the difference in electrostatic free energies of the complex and of the free protein subunits as follows:

$$\Delta\Delta G_{\mathrm{el}} = \Delta G_{\mathrm{el}}(\mathrm{A:B}) - \Delta G_{\mathrm{el}}(\mathrm{A}) - \Delta G_{\mathrm{el}}\ (\mathrm{B}) \qquad (2)$$

where $\Delta G_{\mathrm{el}}$(A:B), $\Delta G_{\mathrm{el}}$(A), and $\Delta G_{\mathrm{el}}$(B) are the electrostatic energies of the complex AB, monomer A, and monomer B as described previously.[51] ICM software[52] was used to calculate the electrostatic free energies of the complex and the free subunits, as DCOMPLEX does not provide a breakup of the binding energy into components. The ICM REBEL (Rapid Exact-Boundary Electrostatics)[53] module included in the ICM-Pro package uses 'boundary element' method to solve Poisson equation for the protein with analytical molecular surface as dielectric boundary. The ICM method is fast and accurate and estimates the electrostatic energy of proteins surrounded by continuous aqueous solution. The energy solved by this method consists of solvation energy and coulomb energy.

### Statistical analysis

The Kruskal-Wallis rank sum test, a nonparametric method,[54] is used to test whether the mean ranks for the PPI features in all functional groups are the same. The discriminatory PPI features among functional groups were thus tested for statistical significance with $P < 0.05$ (for the Kruskal-Wallis rank sum test) in RStudio.[55] Spearman's rank correlation coefficient has been used to measure of statistical dependence between the interface features, since it assesses how well the relationship between two properties can be described using a monotonic function.[56]

## Acknowledgments

G.S. is the recipient of an iMQRES PhD scholarship from Macquarie University. The authors are grateful to Prof Pandjassarame Kangueane and Prof. Michael Gromiha for valuable discussions.

## References

1. Sowmya G, Ranganathan S (2014) Protein-protein interactions and prediction: a comprehensive overview. Protein Pept Lett 21:779–789.
2. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G (2007) The molecular architecture of protein-protein binding sites. Curr Opin Struct Biol 17: 67–76.
3. Keskin O, Gursoy A, Ma B, Nussinov R (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev 108: 1225–1244.
4. Sudha G, Nussinov R, Srinivasan N (2014) An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. Prog Biophys Mol Biol 116:141–150.
5. Bahadur RP, Zacharias M (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cell Mol Life Sci 65:1059–1072.
6. Piehler J (2005) New methodologies for measuring protein interactions in vivo and in vitro. Curr Opin Struct Biol 15:4–14.
7. Sowmya G, Khan JM, Anand S, Ahn SB, Baker MS, Ranganathan S (2014) A site for direct integrin αvβ6.uPAR interaction from structural modelling and docking. J Struct Biol 185:327–335.
8. Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19 Suppl 1:i197–204.
9. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 (Suppl 1):i302–i310.
10. Wiederstein M, Gruber M, Frank K, Melo F, Sippl MJ (2014) Structure-based characterization of multiprotein complexes. Structure 22:1063–1070.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242.

12. Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63:31–65.
13. Gromiha MM, Yokota K, Fukui K (2009) Energy based approach for understanding the recognition mechanism in protein-protein complexes. Mol Biosyst 5:1779–1786.
14. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci USA 93:13–20.
15. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36:21–42.
16. Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256:705–708.
17. Miller S, Lesk AM, Janin J, Chothia C (1987) The accessible surface area and stability of oligomeric proteins. Nature 328:834–836.
18. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. Proteins 53:708–719.
19. Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol 403:660–670.
20. McCoy AJ, Chandana Epa V, Colman PM (1997) Electrostatic complementarity at protein/protein interfaces. J Mol Biol 268:570–584.
21. Xu D, Tsai CJ, Nussinov R (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 10:999–1012.
22. Kundrotas PJ, Alexov E (2006) Electrostatic properties of protein-protein complexes. Biophys J 91:1724–1736.
23. Wong ET, Na D, Gsponer J (2013) On the importance of polar interactions for complexes containing intrinsically disordered proteins. PLoS Comput Biol 9:e1003192.
24. Saha RP, Bahadur RP, Pal A, Mandal S, Chakrabarti P (2006) ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces. BMC Struct Biol 6:11
25. Dong F, Zhou HX (2006) Electrostatic contribution to the binding stability of protein-protein complexes. Proteins 65:87–102.
26. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285:2177–2198.
27. Guharoy M, Chakrabarti P (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics 11:286.
28. Nooren IM, Thornton JM (2003) Diversity of protein-protein interactions. Embo J 22:3486–3492.
29. Park SH, Reyes JA, Gilbert DR, Kim JW, Kim S (2009) Prediction of protein-protein interaction types using association rule based classification. BMC Bioinformatics 10:36.
30. Nooren IM, Thornton JM (2003) Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 325:991–1018.
31. Ozbabacan SE, Engin HB, Gursoy A, Keskin O (2011) Transient protein-protein interactions. Protein Eng Des Sel 24:635–648.
32. Sowmya G, Anita S, Kangueane P (2011) Insights from the structural analysis of protein heterodimer interfaces. Bioinformation 6:137–143.
33. Yugandhar K, Gromiha MM (2014) Protein-protein binding affinity prediction from amino acid sequence. Bioinformatics 30:3583–3589.
34. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 336:943–955.
35. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202.
36. Zhanhua C, Gan JG, Lei L, Mathura VS, Sakharkar MK, Kangueane P (2005) Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in euclidian space. Front Biosci 10:844–852.
37. Zhanhua C, Gan JG, Lei L, Sakharkar MK, Kangueane P (2005) Protein subunit interfaces: heterodimers versus homodimers. Bioinformation 1:28–39.
38. Chen J, Sawyer N, Regan L (2013) Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. Protein Sci 22:510–515.
39. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.
40. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68.
41. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Pérez-Cano L, Pons C, Fernández-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodríguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D. (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414:289–302.
42. Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379–400.
43. Tsodikov OV, Record MT Jr., Sergeev YV (2002) Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J Comput Chem 23:600–609.
44. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. Proteins 47:334–343.
45. Hausman REC, Geoffrey M (2004) The cell: a molecular approach. Washington, DC: ASM Press.
46. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793.
47. Gupta PS, Mondal S, Mondal B, Islam RN, Banerjee S, Bandyopadhyay AK (2014) SBION: A program for analyses of salt-bridges from multiple structure files. Bioinformation 10:164–166.
48. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774–797.
49. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 56:93–101.
50. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11:2714–2726.

51. Talley K, Ng C, Shoppell M, Kundrotas P, Alexov E (2008) On the electrostatic component of protein-protein binding free energy. PMC Biophys 1:2.

52. Abagyan RA, Totrov MM, Kuznetsov DA (1994) ICM: A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comp Chem 15:488–506.

53. Abagyan RA, Totrov MM (2001) Rapid boundary element solvation electrostatics calculations in folding simulations: successful folding of a 23-residue peptide. Biopolymers 60:124–133.

54. McDonald JH (2014) Handbook of biological statistics. Baltimore, MD: Sparky House Publishing.

55. RStudio. RStudio: Integrated development environment for R v0.98.1091. Boston 2012.

56. Corder GW, Foreman DI (2014) Nonparametric statistics: a step-by-step approach. New York: Wiley.

## SUPPLEMENTARY MATERIAL

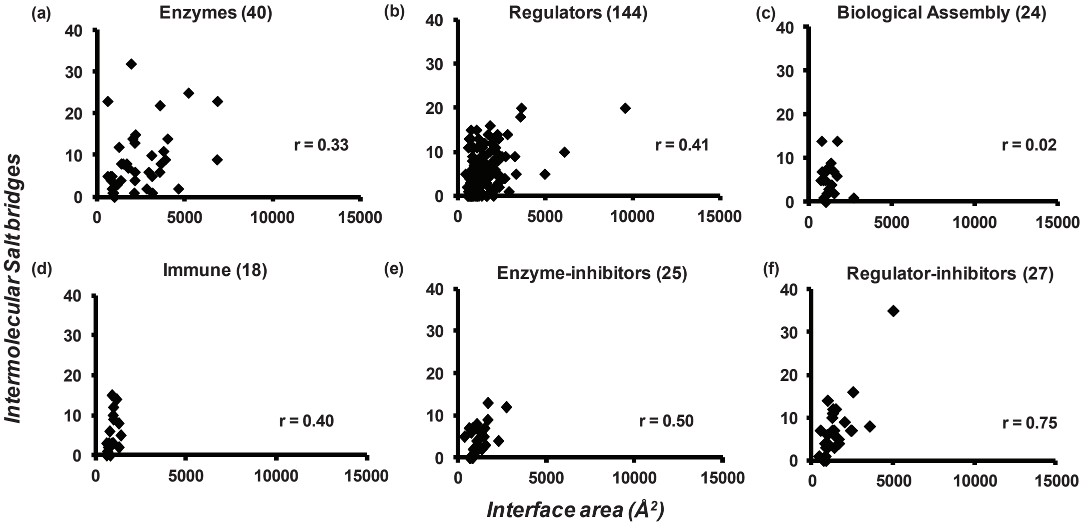


**Figure S1: Intermolecular salt bridges show limited correlation with interface area.** Intermolecular salt bridges at the protein interfaces are poorly correlated to interface area in most functional classes, however (f) regulator-inhibitors show high correlation (r = 0.75).

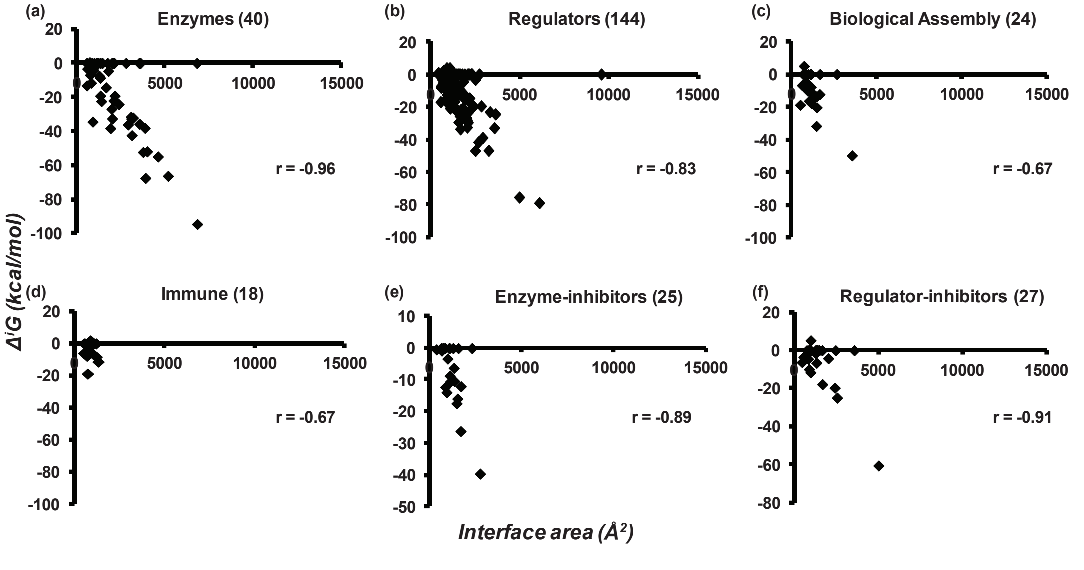


**Figure S2: Solvation-free-energy-gain upon interface formation shows high correlation with interface area.** $\Delta^{i}G$ shows high correlation with interface area in (a) enzymes (r = -0.96), (b) regulators (r = -0.83), (e) enzyme-inhibitors (r = -0.89), and (f) regulator-inhibitors (r = -0.91), however, (c) biological assembly and (d) immune complexes show relatively less correlation with interface (r = -0.67) as compared to other functional classes.

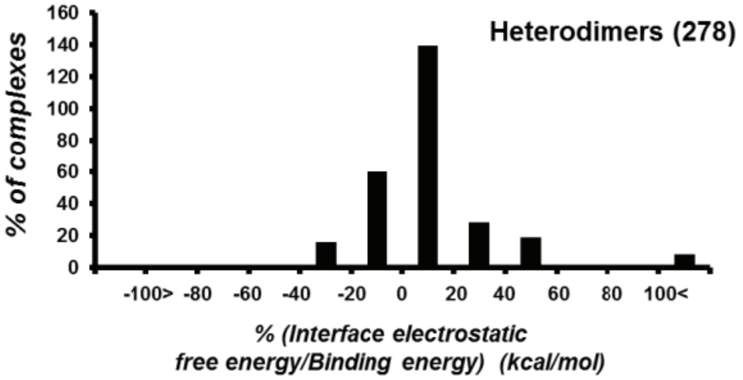


**Figure S3: Percentage interface electrostatic free energy component of binding energy distribution.** The quantitative distribution of interface electrostatic energy contributes to destabilizing PPIs in a few complexes while stabilizing PPIs in the others.

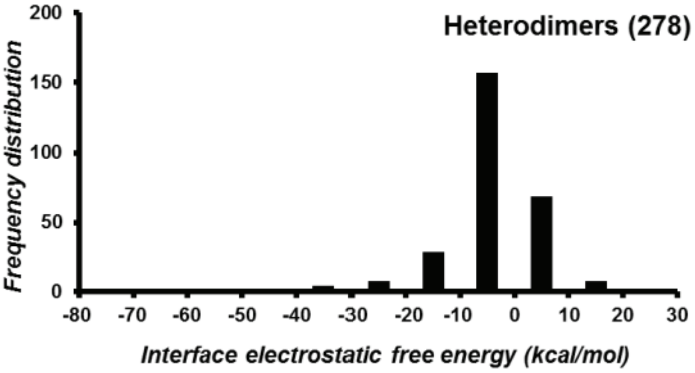
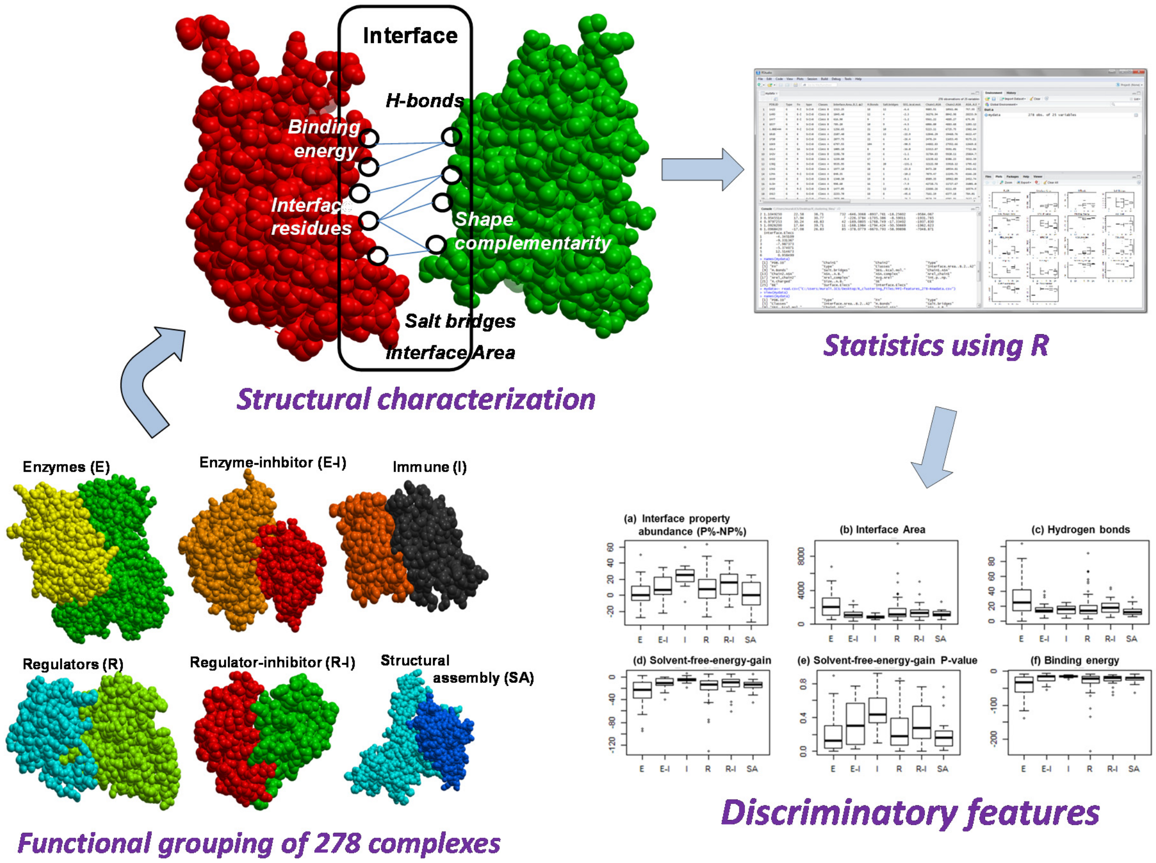


**Figure S4: Interface electrostatic energy frequency distribution.** The frequency distribution shows similar trends in interface electrostatic energy.

**Table S1: Discriminatory PPI features among functional groups in globular protein types (using Kruskal-Wallis rank sum test)**

| Functional groups | P-value | Q-value |
|---|---|---|
| ***Solvent-free-energy-gain ($\Delta^i G$)*** | 3.12E-05 | 0.000561 |
| ***Interface Area*** | 9.89E-05 | 0.001681 |
| ***Binding Energy*** | 1.62E-04 | 0.002427 |
| ***Interface property abundance (P%-NP%)*** | 1.22E-04 | 0.001945 |
| ***Hydrogen Bonds*** | 3.23E-03 | 0.041998 |

# GRAPHICAL ABSTRACT



## 4.3 Conclusions

Knowledge of the molecular principles of protein-protein binding is essential for the understanding of complex mechanisms in cellular processes. Therefore, it is of interest to analyse protein interfaces of known complexes for functional interpretation. Interfaces are parts of protein surfaces in the unbound state whose differential physico-chemical properties compared to the rest of the complex's surface drive binding and complex formation. Hence, a comprehensive analysis of known structural interfaces of 278 heterodimer complexes was completed and their gleaned interface features are documented in this study.

Although, functional ascription of structural interface is a non-trivial task, key interface properties are able to distinguish between functional classes of complexes. I have documented key discriminatory features (interface area, interface property abundance (P%-NP%), H-bonds, $\Delta^{i}G$ and BE) in different functional classes of complexes. These representative features have implications for the prediction of potential protein function of novel complexes. The results provide molecular insights for better understanding of PPIs and their relation to biological functions. With the availability of additional information, such as dynamics for these complexes, it will be possible to comment on the dynamic stability of the dataset, in the future.

This study lead to the preliminary application of using PPI analysis for the identification of the integrin αvβ6·uPAR interactions which are crucial for cancer progression (detailed in chapter 5).

# Chapter 5: Dissecting interfaces of interacting proteins: integrin αvβ6·uPAR interactions

## 5.1 Summary (Preliminary application)

The integrin αvβ6 heterodimer was studied for the characterisation of protein-protein interfaces, using information obtained from the previous studies.

A comprehensive analysis of the PPIs involving integrin αvβ6 and its binding partners helps in better understanding the structural basis of integrin activation. Moreover, there is increasing evidence primarily from *in vitro* studies showing that the integrin αvβ6•uPAR interactions play a crucial role in cancer progression. However, the complete 3D X-ray crystallographic structure of integrin αvβ6 heterodimer remains elusive possibly due to its large size, membranous nature and complexity. Moreover, experimental determination of the 3D structure of large membrane proteins remains a laborious task; therefore homology modeling procedures are indispensable tools for structure prediction and structure based drug designing [323].

Composite homology modeling approaches to build the complete 3D structural model of integrin αvβ6, including the transmembrane and cytoplasmic regions of the two subunits using other known integrin X-ray structures as templates is employed. Subsequently, structural analysis (detailed in Section 1.4.3.1) of integrin αvβ6•uPAR interactions was performed using model data with docking simulation for their binding. The interaction region and site on domain III of uPAR and αv subunit is in consensus with experimental data (detailed in Appendix 1 – Publication 5) providing high-affinity potential sites of interaction in 3D space.

The molecular basis of integrin αvβ6•uPAR binding using structural data is discussed (in Publication 4 [324]) for implications as potential therapeutic targets in cancer management.

## *5.2 Publication 4*

Journal of Structural Biology 185 (2014) 327–335

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

# A site for direct integrin αvβ6·uPAR interaction from structural modelling and docking

Gopichandran Sowmya [a,b], Javed Mohammed Khan [a,b], Samyuktha Anand [a], Seong Beom Ahn [a], Mark S. Baker [a], Shoba Ranganathan [a,b,c,]*

[a] Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia
[b] ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia
[c] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

ARTICLE INFO

*Article history:*
Received 16 September 2013
Received in revised form 22 November 2013
Accepted 7 January 2014
Available online 11 January 2014

*Keywords:*
Integrin heterodimer
Integrin αvβ6
uPAR
Structural modelling
Protein–protein interactions
Molecular docking

ABSTRACT

Integrin αvβ6 is an epithelially-restricted heterodimeric transmembrane glycoprotein, known to interact with the urokinase plasminogen activating receptor (uPAR), playing a critical role in cancer progression. While the X-ray crystallographic structures of segments of other integrin heterodimers are known, there is no structural information for the complete αvβ6 integrin to assess its direct interaction with uPAR. We have performed structural analysis of αvβ6·uPAR interactions using model data with docking simulations to pinpoint their interface, in accord with earlier reports of the β-propeller region of integrin α-chain interacting with uPAR. Interaction of αvβ6·uPAR was demonstrated by our previous study using immunoprecipitation coupled with proteomic analysis by mass spectrometry. Recently this interaction was validated with proximity ligation assays and peptide arrays. The data suggested that two potential peptide regions from domain II and one peptide region from domain III of uPAR, interact with αvβ6 integrin. Only the peptide region from domain III is consistent with the three-dimensional interaction site proposed in this study. The molecular basis of integrin αvβ6·uPAR binding using structural data is discussed for its implications as a potential therapeutic target in cancer management.

## 1. Introduction

Integrins are a large family of αβ heterodimeric cell surface receptors found in multicellular organisms, from sponges to mammals (Humphries, 2000; Kim et al., 2011). These receptors are involved in bidirectional cell signalling, transducing information between the components of extracellular and intracellular milieu (Hynes, 2004). Binding of integrin glycoproteins to the extracellular matrix initiates structural changes, consequently triggering signal transduction (Xiong et al., 2001, 2002). These signals, arising from receptor-binding, are implicated in cell migration, attachment, differentiation, proliferation, polarity and survival/apoptosis (Giancotti and Ruoslahti, 1999; Xiong et al., 2001). Involvement of integrin receptors in initiation and/or progression of many malicious diseases including tumour metastasis, immune dysfunction, neoplasia, inflammation, trauma and infections have been reviewed elsewhere (Arnaout et al., 2007; Hynes, 2002, 2004). Integrins are also receptors for many viruses and bacteria, and have been the target of therapeutic drugs to combat inflammation, thrombosis, fibrosis and tumourigenesis (Binder and Trepel, 2009; Hynes, 2002; Lu et al., 2008; Margadant and Sonnenberg, 2010; Van Aarsen et al., 2008).

Integrins consist of two distinct subunits (α and β) forming heterodimers with an obligatory function (constituent subunits are unstable in monomeric form). These α and β subunits assemble into a "head" segment built on top of two V-shaped 'legs' (Campbell and Humphries, 2011). In mammals, integrins noncovalently associate to form 24 different receptors assembled from 18 α subunits and 8 β subunits (Barczyk et al., 2010; Hynes, 2002). Each of these 24 receptors has been characterised to have a unique function based on their ligand-binding specificities. The R–G–D (arginine–glycine–aspartate; RGD) tri-peptide sequence is a commonly

*Abbreviations:* 3D, three-dimension; aa, amino acid residues; ASA, accessible surface area; EMT, epithelial to mesenchymal transition; ERK2, extracellular signal-regulated kinase; GPI, glycosylphosphotidylinositol; ICM, internal coordinate mechanics; LAP, latency associated peptide; MMP, matrix metalloproteinase production; MSEP, molecular surface electrostatic potential; PIPSA, protein interaction property similarity analysis; PPI, protein–protein interaction; PSVS, protein structure validation suite; RGD, arginine–glycine–aspartate tripeptide; RMSD, root mean square deviation; SAVES, the structure analysis and verification server; TGFβ, transforming growth factor-beta; uPA, urokinase plasminogen activator; uPAR, urokinase plasminogen activating receptor; VN, vitronection.

* Corresponding author at: Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia. Fax: +61 2 9850 8313.
*E-mail address:* shoba.ranganathan@mq.edu.au (S. Ranganathan).

1047-8477/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2014.01.001

known integrin-binding motif associated with the αβ interface of several integrins, although individual integrins specifically bind to select protein ligands (Takada et al., 2007). Each subunit has a comparatively large extracellular domain, a transmembrane domain and a short cytoplasmic tail (Hynes, 2002; Kim et al., 2011). The sequence lengths of α and β subunits of an integrin heterodimer are at least 1000 and 750 amino acid (aa) residues respectively. In the past decade, elucidation of three-dimensional (3D) structures of specific domains of integrins (Xiong et al., 2002, 2009) has paved the way for researchers to perform intensive structural analyses related to the functional significance of these large glycoproteins.

Integrin αvβ6 is a member of the integrin family, with a β6 subunit which binds exclusively to the αv subunit (Bandyopadhyay and Raghavan, 2009; Busk et al., 1992). Expression of αvβ6 integrin is primarily restricted to epithelial cells, where it is expressed at low-levels in normal adult cells and is elevated during embryogenesis, morphogenesis (i.e., epithelial to mesenchymal transition, EMT), injury, wound healing and tumourigenesis (Bandyopadhyay and Raghavan, 2009; Bates et al., 2005; Breuss et al., 1995). αvβ6 integrin mediates cell division, adhesion, migration and invasion. It is also known to contribute to the complex mechanism of EMT, from initiating cell signalling cascades and interacting with other membrane proteins to activating extracellular signals such as latent transforming growth factor β (TGF-β), a recognised inducer of EMT. The interaction with TGF-β is through the RGD motif present in the latency associated peptides 1 and 3 (LAP1 and LAP3) (Busk et al., 1992; Morris et al., 2003; Munger et al., 1999). The β6 subunit has a unique C-terminal cytoplasmic sequence that directly interacts with the extracellular signal-regulated kinase (ERK2) to activate the ERK/MAPK pathway, often highly activated during tumour progression and metastasis, and to produce matrix metalloproteinase production (MMP2 or MMP9) (Morgan et al., 2004; Bandyopadhyay and Raghavan, 2009). Binding of integrin αvβ6 to the glycosylphosphotidylinositol (GPI) anchored membrane protein urokinase plasminogen activating receptor (uPAR) promotes the plasminogen activator system (Saldanha et al., 2007).

uPAR is a versatile signalling orchestrator mediating interactions with other transmembrane receptors, including integrins. The uPAR comprises three domains which are anchored by glycosylphosphotidylinositol (GPI) to the extracellular surface of the plasma membrane. These three domains form a glove-like structure that provides a central pocket for the binding of uPA (Smith and Marshall, 2010). uPAR is believed to play a role in downstream cellular signalling pathways through lateral interactions with transmembrane proteins such as integrins as they lack intrinsic intracellular domains (Eden et al., 2011). Therefore, an in-depth understanding of the protein–protein interactions involving integrin αvβ6 and uPAR, is of immense interest. However, the complete X-ray crystallographic structure of integrin αvβ6 heterodimer remains elusive.

Experimental determination of the 3D structure of large membrane proteins remains a laborious task; therefore homology modelling procedures are indispensable tools for structure prediction and structure based drug designing (Sanchez and Sali, 1997). We have employed composite homology modelling approaches to build the complete 3D structural model of integrin αvβ6, including the transmembrane and cytoplasmic regions of the two subunits. Analysis of the structural properties of integrin αvβ6 in comparison with other X-ray crystal structures of integrins known to interact with uPAR reveals that αvβ6 integrin is independently an RGD-binding and a uPAR-binding receptor. Docking simulations between the integrin αvβ6 structural model and uPAR protein reveal a single potential interaction site, thereby providing better understanding of integrin αvβ6 mediated regulation of the plasminogen activator system and also help gain insights into integrin αvβ6·uPAR interactions.

## 2. Materials and methods

Protein structures are evolutionarily more conserved when compared to their sequence conservation alone (Sander and Schneider, 1991). This observation led to the development of homology modelling, a powerful technique which has been used for many years to computationally build 3D structural models of proteins, provided a structural homologue with more than 20% sequence similarity is available (Huynh et al., 2011; Khan and Ranganathan, 2009; Ranganathan, 2001). Homology modelling provides reliable and qualitative structural models for further investigations into the structure–function relationship of a protein and also, in many cases, directs further experimental studies. Moreover, it is well known that integrins are a family of proteins with a highly conserved overall structure. Therefore, we have applied homology modelling to build a comprehensive high-quality 3D structural model of integrin αvβ6 and obtain crucial knowledge of the vital integrin αvβ6·uPAR interactions which have known implications in cancer metastasis.

### 2.1. Data collection

Complete sequences of human integrin αv (UniProt: P06756 with a chain length of 1048 aa) and β6 (UniProt: P18564 with a chain length of 788 aa) subunits were retrieved from the UniProtKB/SwissprotKB database release (2014). Currently, there is no complete structure (encompassing the extracellular, transmembrane and cytoplasmic domains) for any αβ integrin heterodimer in the Protein DataBank (PDB) (Berman et al., 2000). Nevertheless, X-ray crystal structures for complete extracellular domains and short C-terminal transmembrane stretches of integrin αvβ3 (PDB: 3IJE) (Xiong et al., 2009) with a query coverage of 96% (identities: 347/693 (50%), positives: 469/693 (68%), gaps: 11/693 (2%) between β3 and β6) was identified as the closest homologue to β6 sequence (obtained by performing a NCBI BLAST (Altschul et al., 1997), search against the PDB protein sequences). The integrin αv chain was derived from the αvβ3 structure for modelling. The 3IJE structure of αvβ3 (the template structure used for modelling αvβ6) was resolved in a bent conformation under conditions that activate ligand binding in biochemical and cell biological assays (Xiong et al., 2009). The transmembrane-cytoplasmic domains of integrin αIIbβ3 (PDB: 2KNC) (Yang et al., 2009) was the closest homologue, with a query coverage of 90% (identities: 10/14 (71%), positives: 10/14 (71%), gaps: 2/14 (14%)) and 77% (identities: 29/45 (64%), positives: 37/45 (82%), gaps: 0/45) for transmembrane and cytoplasmic domains of αv and β6 respectively. Hence, these X-ray crystal structures were chosen as the two templates for 3D structural modelling of the integrin αvβ6 complex. Sequences of the template structures were extrapolated from their respective X-ray crystal structures using MODELLER (Sali and Blundell, 1993).

### 2.2. Preliminary evaluation

Alignment of the target sequence with that of the template structure is the most critical step in modelling an accurate structure. The retrieved sequences of integrin αv and β6 subunits were aligned to the extracted sequences from the template structures of integrin αvβ3 (PDB: 3IJE; for extracellular domains) and integrin αIIbβ3 (PDB: 2KNC; for transmembrane-cytoplasmic domains), using ClustalX (Thompson et al., 1997) with default BLOSUM scoring matrices. Prior to the alignment process, we noted that UniProt has annotated the presence of signal peptides in integrin αv (30 aa) and β6 (22 aa) sequences. These were removed from the respective protein sequences as they are absent in the mature template protein structures. Moreover, due to the absence of a few residues in the integrin αvβ3 (PDB: 3IJE) template structure, the gaps in the alignment were carefully scrutinised and manually curated to retain chain boundaries and the conservation of structurally and functionally important residues (Supplementary Fig. S1). Furthermore, in view of the fact that there were multiple templates being used for the modelling, we allowed an overlap of at least four residues between the template sequences (the complete alignment is provided in Supplementary Fig. S1, annotated with the secondary structure of the template structures), to ensure that the relevant biological orientation is adopted. The alignment obtained has seven residues overlapping between the two template structures used for both modelling integrin αv and β6 sequences, as shown in Supplementary Fig. S1.

### 2.3. 3D structural modelling, refinement and quality verification

The model building process was carried out using MODELLER (version 7v7) (Sali and Blundell, 1993), owing to its superiority over other homology modelling software in allowing the simultaneous generation of a multi-chain protein assembly, ligand inclusion and the use of multiple templates for model building (Huynh et al., 2011; Khan and Ranganathan, 2009; Tng et al., 2004). The optimal satisfaction of dihedral angle restraints and spatial constraints employed in MODELLER, ensure sound stereochemistry for the structural models.

We have initially generated a structural model without constraints for disulfide bonds or calcium ions (data not shown). We found that six disulphide bonds (one in the αv subunit and five in the β6 subunit) annotated by Uniprot/Swiss-Prot were missing in the initial model. The residue numbers for the missing disulfide bonds were specifically constrained in the model building command file so that all missing disulfide bonds were included in our structural models. In addition, the existence of four ligands ($Ca^{2+}$ ions), all binding to αv subunit, has been annotated in UniProt, whereas the template structure of integrin αvβ3 contains six ligands ($Ca^{2+}$ ions; five bound to αv subunit and one bound to β3 subunit). As all six $Ca^{2+}$ ion binding sites are viable from sequence conservation, we have retained all six ligands in our integrin αvβ6 models by extrapolating (and refining) their structural coordinates and positional information from the integrin αvβ3 template structure, in order to dissect the effect of ligand binding on both αv and β6 subunit interactions and integrin αvβ6·uPAR interactions.

Five composite models of integrin αvβ6 heterodimeric complex were generated based on the sequence alignment and the best model was selected on the basis of MODELLER's DOPE score.

Upon structural refinement, three models with the best objective functions, out of the five, were chosen. Subsequently, the model with the lowest current energy among these three was selected based on stereo-chemical quality assessment. The Protein Structure Validation Suite (PSVS) (Bhattacharya et al., 2007) and the Structure Analysis and Verification Server (SAVES; http://services.mbi.ucla.edu/SAVES/) version 4, which incorporate major structural assessment tools such as PROCHECK (Laskowski et al., 1996), PDB Validation (Diago et al., 2007) and WHATCHECK (Hooft et al., 1996), were used to evaluate the overall accuracy of the structural model by assessing residue geometry, protein folding, bond-length, bond-angle, stereo-chemical quality, possible errors in localised regions and performing volumetric analysis. Internal Coordinate Mechanics (ICM) package version 3.7-2a (Abagyan et al., 1994) was used to visualise the structures and PyMOL was utilised to calculate the root mean square deviation (RMSD) values of superimposed Cα positions. After performing stereo-chemical quality checks and structural assessment, the best quality integrin αvβ6 heterodimer model was selected for further structural analyses. The coordinates of the final model in PDB format are available from Supplementary Fig. S2.

### 2.4. Structural analyses

#### 2.4.1. Interface and surface analysis

Interface residues in a protein–protein complex are identified by calculating the interface area or change in solvent accessible surface area (ΔASA) upon complex formation. Similarly, surface residues are identified by calculating the ASA of the protein complex. The efficiency of Surface Racer 5.0 (Tsodikov et al., 2002) to accurately calculate ASA and determine interacting residues in a protein complex has been previously documented (Sowmya et al., 2011). Therefore, we used Surface Racer (Tsodikov et al., 2002) to identify both the interface residues between integrin αv and β6 subunits and the solvent exposed surface residues of integrin αvβ6 complex. ΔASA of integrin αvβ6 complex was calculated as follows:

$$\Delta ASA(\alpha v\beta 6) = [ASA(\alpha v) + ASA(\beta 6) - ASA(\alpha v\beta 6)]/2 \quad (1)$$

A probe radius of 1.4 Å was used to calculate ASA (Lee and Richards, 1971). The interface residues were further filtered based on the criteria that their relative ΔASA is at least 4% and not less than 5 $Å^2$ upon complex formation (Porollo and Meller, 2007).

The residue compositions of interface and surface residues of integrin αvβ6 complex were documented in order to determine the type of interactions (obligatory, occurring at the interface or transient, occurring at the surface) of the integrin αvβ6 heterodimer. The residues are grouped as polar and non-polar, based on their residue type. We then estimated the amount of polar, nonpolar and charged residues at the surface and the interface of integrin αvβ6 complex. The residue composition for surface residues of uPAR was also identified to determine the type of surface interactions that occur abundantly on uPAR.

#### 2.4.2. Molecular surface electrostatic potential (MSEP) calculation and comparison

A result of charged side chains of the amino acid residues and bound ions, MSEP in proteins plays a vital role in protein folding, stability, enzyme catalysis and specific protein–protein interactions. The measure of similarity in the composition of charged residues between any two (or a group of) proteins is their MSEP similarity. We calculated and compared the electrostatic interaction properties of the ectodomains of our integrin αvβ6 model with all available crystal structures of αβ integrins (five; four uPAR-binding heterodimers {αvβ3, PDB: 3IJE; αIIbβ3, PDB: 3FCS (Zhu et al., 2008); α5β1, PDB: 3VI3 (Nagae et al., 2012); αXβ2, PDB: 3K6S (Xie et al., 2010)} and one uPAR non-binding dimer {α4β7, PDB: 3V4V (Yu et al., 2012)} as a control) using the webPIPSA (Richter et al., 2008) server of the Protein Interaction Property Similarity Analysis (PIPSA; Blomberg et al., 1999) program.

Initially, the algorithm calculates MSEP for all proteins. Similarity indices are then calculated for all pairs of proteins based on the electrostatic similarity. Subsequently, electrostatic distances are computed using the similarity indices. These electrostatic distances are then plotted as a tree/cluster dendogram and as a colour coded matrix called a heat map. These cluster dendograms and heat maps were consequently used to study and understand the similarities between uPAR binding αβ integrins and the differences between uPAR binding and non-uPAR binding αβ integrins.

### 2.5. Docking of integrin αvβ6 model with uPAR X-ray crystal structure

The ICM (Abagyan et al., 1994) package is benchmarked and extensively validated for docking simulations and ligand binding mode accuracy (Neves et al., 2012), for flexible protein–protein interaction modelling. Hence, we used ICM for our docking project to generate multiple conformations, owing to its high success rates in including experimentally solved near-native conformations. The ICM protein–protein docking protocol was used to perform docking simulations between our integrin αvβ6 model and the crystal structure of uPAR (PDB: 1YWH; 268aa, residues renumbered as per UniProt: Q03405).

Docking is an extremely compute-intensive procedure, therefore the complete structural model of integrin αvβ6 was not considered suitable for docking simulations. Experimental reports on the interactions of αβ integrins with uPAR provide evidence that the β-propeller region of integrin α-chain is involved in direct interactions with uPAR affecting the functions of αβ integrins (Chaurasia et al., 2006; Simon et al., 2000; Zhang et al., 2003). Hence, we sliced our complete integrin αvβ6 model for β-propeller domain of integrin α-chain (440 aa) and its corresponding binding domain in the β-chain (244 aa) to carry our docking simulations with uPAR crystal structure. This 684 aa αvβ6 β-propeller region was considered as a "receptor" and uPAR (268 aa) was considered as a "ligand" for docking simulations.

The uPAR binds uPA at an internal cavity, leaving the large outer surface of uPAR for interactions with other ligands such as integrins and vitronectins (Llinas et al., 2005). Hence, the outer surface residues of uPAR were selected as epitopes or potential binding sites of interest in our docking simulations, with complete freedom to bind anywhere to the integrin αvβ6 β-propeller region. Three docking simulation projects were set up by selecting different 'epitopes' on the ligand, uPAR (the outer surface residues of domain I, II and III, the outer surface of domain III only and the binding sites from previous studies (Chaurasia et al., 2006), to avoid bias in the selection of potential uPAR binding sites and also to optimise and verify interaction sites in docked complexes. Docking refinements using the ICM global optimisation docking algorithm were then performed for complexes, to minimise conformational energies in all the three docking projects. Subsequently, αvβ6–uPAR interaction analysis was performed on the three refined complexes generated from the docking projects. The ICM docking project (20 *in silico* experiments) was run on a 2 CPU 2.66 GHz 24 GB RAM workstation, with each run taking about 48 h.

## 3. Results and discussion

### 3.1. High-quality 3D structural model for integrin αvβ6

The 3D structural models generated by MODELLER are internally subjected to multiple iterations of stereo-chemical refinements upon their selection from a pool of randomized potential starting conformations. Moreover, the structural models that MODELLER generates also satisfy spatial restraints, *de novo* loop modelling along with structure optimisation with respect to flexibly defined objective function. The structural quality assessment of the best integrin αvβ6 model (Fig. 1) (selected based on the best objective function and lowest current energy as explained above) was performed using PSVS and SAVES structure validation servers as described earlier.

The PROCHECK program embedded in the PSVS package calculated that our integrin αvβ6 model had 97.2% of residues within the conformationally allowed regions, while the minimum benchmark for a high quality X-ray crystal structure is 85%. These comprise 84.0% in the most favoured and 13.2% in additional favoured regions. Moreover, the PDB Validation software (also embedded in the PSVS package) calculated an average value of 2.5° and 0.020 Å for deviations from ideal geometry for bond angles (BA) and bond lengths (BL), respectively, which is typical for a good resolution X-ray crystal structure (Diago et al., 2007).

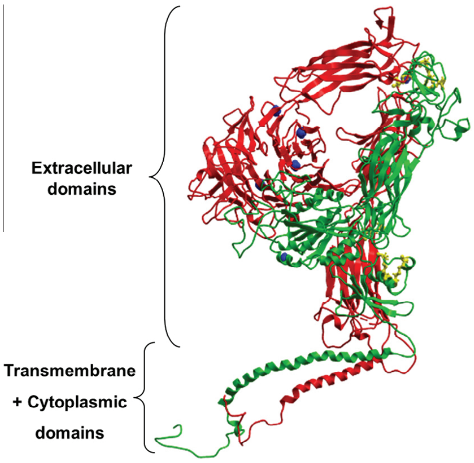


**Fig.1.** The complete structural model of integrin αvβ6 in cartoon representation. αv and β6 subunits are shown in red and green, respectively. The six ligands ($Ca^{2+}$ ions) are depicted as blue spheres while the newly built disulfide bonds (yellow) are in ball and stick representation. The extracellular and the transmembrane and cytoplasmic domains of the two subunits are labelled. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The WHATCHECK program (included in the SAVES package) checks normality of the local environment of amino acids to estimate the packing quality of a 3D protein structure. Our model returned a First Generation Packing Quality (FGPQ) *Z*-score of −1.580, which lies within the range of −2 to 0 for a good structural model. WHATCHECK also computed the RMS *Z*-scores of 1.301 and 0.942 for bond angles and bond lengths, respectively. The close proximity of these values to 1.5 for bond angles and 1.0 for bond lengths (RMS *Z*-score for an ideal structure) signifies a good structural model. Moreover, the model shares a high degree of conservation of functional domains with the αvβ3 (PDB: 3IJE) template structure. The RMSD calculations performed using PyMOL retrieved scores of 0.853 and 0.500 Å for the ectodomains of αv and β6 subunits, respectively (model compared to template structure of αvβ3, PDB: 3IJE). Similarly, the RMSD values for transmembrane and cytoplasmic domains of αv and β6subunits (model compared to template structure of αIIbβ3, PDB: 2KNC) are 1.493 and 1.071 Å, respectively. The model was generated for αv and β6 chains together as a heterodimeric structure. Two templates were used to build the αv chain (ectodomain and transmembrane + cytoplasmic regions from αvβ3 and αIIbβ3 structures respectively), with at least four overlapping residues (Supplementary Fig. S1). This overlap (of seven residues) defined the relative orientation of the inter domains, with the available X-ray data. Therefore, the RMSDs of αv chain of αvβ6 model and the templates used (αvβ3 for ectodomain and αIIbβ3 for transmembrane + cytoplasmic regions) were also calculated based on the alignment. Despite the fact that accurate *Z*-scores may not be applicable for the analysis of transmembrane proteins (Bhattacharya et al., 2007), our αvβ6 model has returned excellent *Z*-scores across the structure validation programs used in this investigation suggesting that the αvβ6 structural model is of high quality (Table 1). Hence, this structural model of integrin αvβ6 was utilised for further structural analyses.

**Table 1**
Structure quality assessment of the integrin αvβ6 model.

| PROCHECK | | PDB validation (deviations from ideal geometry) | | WHATCHECK (RMS *Z*-score) | | | RMSD (Å; calculated using PyMOL) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ramachandran plot summary of model residues | | BA (°) | BL (Å) | FGPQ | BA (°) | BL (Å) | Ectodomain (model and 3IJE) | | Transmembrane + Cytoplasmic domains (model and 2KNC) | |
| Most favoured regions (%) | Allowed regions (%) | | | | | | αv | β6 | αv | β6 |
| 84.0 | 15.3 | 2.5 | 0.020 | −1.580 | 1.301 | 0.942 | 0.853 | 0.500 | 1.493 | 1.071 |

### *3.2. Abundance of polar interactions at αvβ6 subunit interface*

The subunit interface analysis of integrin αvβ6 model determined that 194 aa (95 αv aa and 99 β6 aa; 11.05% of the total 1756 aa and 11.97% of the 1621 ectodomain residues) participate in subunit-subunit interactions, forming an obligatory protein–protein interaction complex (i.e. subunits are not stable in their monomeric states and have no independent existence). The interface area formed by the interacting residues of αv and β6 subunits is depicted in Fig. 2a. The integrin αvβ6 complex buries a large interface area of 3725.25 Å$^2$ suggesting that the αv and β6 subunits undergo large conformational dynamics upon assembly (Lo Conte et al., 1999). The interacting residues at the αvβ6 interface are also represented as a function of residue number using ΔASA in an *X*–*Y* plot (Fig. 2b). The hydrophobic core region of the αvβ6 complex comprises of 222 aa (12.64% of the total 1756 aa; 13.70% of the 1621 ectodomain residues), while 1399 aa (79.67% of the total 1756 aa; 86.30% of the 1621 ectodomain residues) are present at its solvent-exposed surface.

The surface residues of αv and β6 subunits that become inaccessible to solvent upon binding and thereby contribute to interface formation were analysed for their chemical nature in order to gain insights into the extent and strength of such protein–protein interactions. Therefore, the chemical properties of interface and surface regions of the integrin αvβ6 were determined based on the contribution of aa residue types. Interestingly, our analysis shows that the two surfaces (subunit interface and the solvent accessible surface) have similar aa compositions. Table 2 shows that the 54.12% polar (27.31% neutral, 11.86% positively charged and 14.95% negatively charged residues) residues occur at the αvβ6 subunit interface, while 45.88% of the interface residues are non-polar (hydrophobic), which is very similar to the αvβ3 subunit interface (55.67% polar and 44.32% non-polar). Similarly, the residues contributing to the surface of the αvβ6 complex are 53.04% polar and 46.96% hydrophobic in nature, with charged residues amounting for 23.30% (positively charged: 10.08% and negatively charged: 13.22%) and 29.74% polar neutral residues. These outcomes

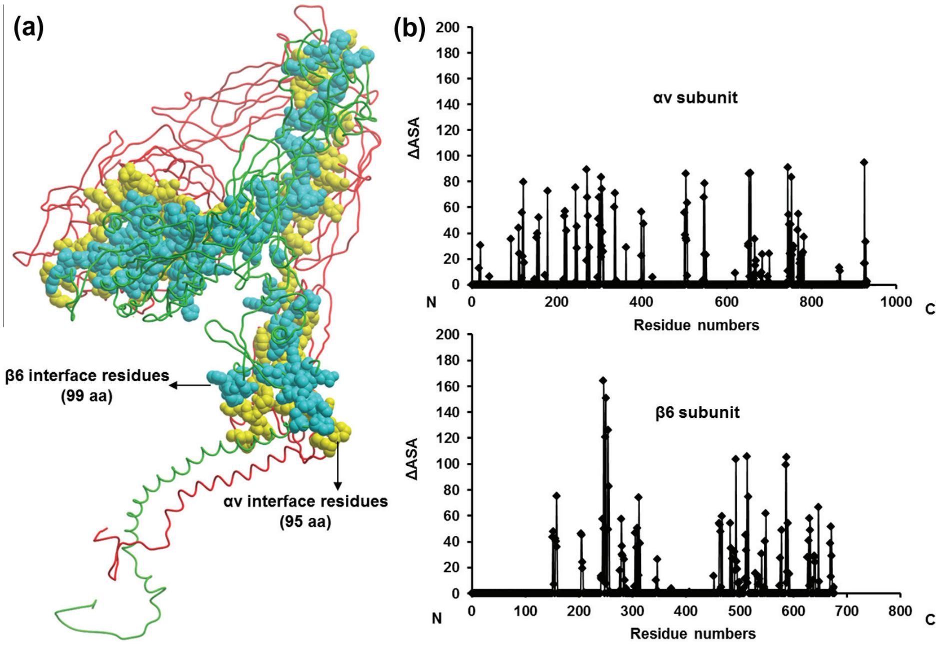


**Fig.2.** (a) Integrin αvβ6 model depicting the interface area formed by the interacting residues of αv and β6 subunits. αv (red) and β6 (green) subunits are shown in cartoon representation. The interface residues from αv (95 aa) and β6 (99 aa) subunits are shown as yellow and cyan spheres, respectively, and are labelled. (b) A graphical representation of the αv and β6 interface residues as a function of residue number using ΔASA. The residue numbers are plotted on the *x*-axis while ΔASA is on the *y*-axis. The N and C termini of the two subunits are portrayed along their corresponding *x*-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Amino acid (aa) residue composition for the solvent accessible surface and the subunit interface of integrin αvβ6 heterodimeric complex. % aa composition is reported for a total of 1621 ectodomain residues.

| aa residue type | | | Surface (% aa composition) | αvβ6 interface (% aa composition) |
|---|---|---|---|---|
| *Polar* | Neutral | | 29.64 | 27.20 |
| | Charged | +ve | 10.40 | 15.40 |
| | | −ve | 13.00 | 11.80 |
| | Total | | 53.04 | 54.40 |
| Non-polar | | | 46.96 | 45.60 |

suggest that polar interactions are the predominant forces that drive both αvβ6 subunit interface formation and protein–protein interactions at the αvβ6 surface.

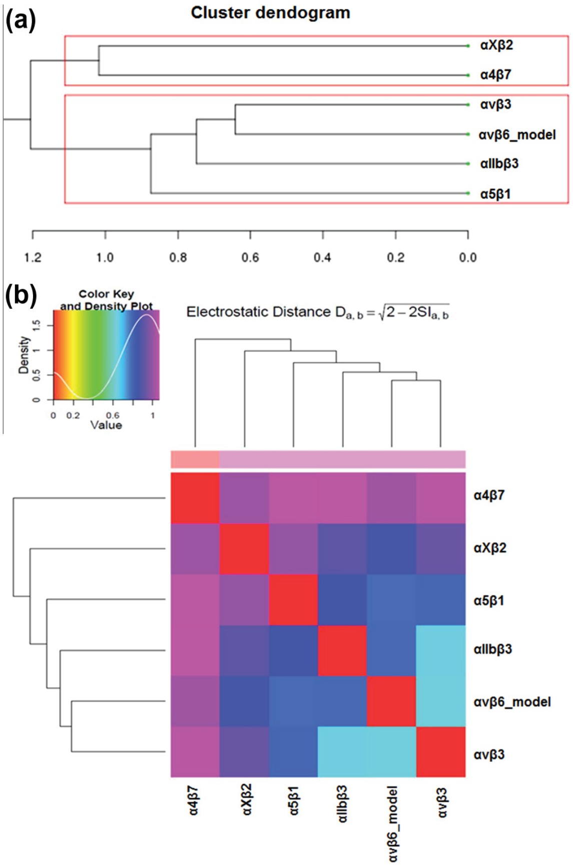


**Fig.3.** (a) Cluster dendogram illustrating the similarity of integrin αvβ6 model to other RGD-binding αβ integrins. The electrostatic surface potentials for the ectodomains of available crystal structures of αβ integrins (αvβ3, PDB: 3IJE; αIIbβ3, PDB: 3FCS; α5β1, PDB: 3VI3 that bind the RGD peptide and α4β7, PDB: 3V4V; αXβ2, PDB: 3K6S that do not bind the RGD peptide) were compared with our αvβ6 model. (b) Heat map depicting the electrostatic similarity of integrin αvβ6 model to other uPAR-binding αβ integrins. Similar to the comparison with RGD binders, our αvβ6 model was also compared to known uPAR-binding αβ integrins (αvβ3, PDB: 3IJE; αIIbβ3, PDB: 3FCS; α5β1, PDB: 3VI3; αXβ2, PDB: 3K6S) and α4β7, PDB: 3V4V (which is not known to bind uPAR and acted as a control). The heat map shows colours ranging from red, indicating most electrostatic similarity to magenta indicating least electrostatic similarity. These MSEP comparisons were carried out using the webPIPSA server (Richter et al., 2008). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *3.3. αvβ6 integrin clusters with other RGD-binding and uPAR-binding receptors*

MSEP similarity calculation of the ectodomains of all available crystal structures of αβ integrins (using the webPIPSA server) confirms that integrin αvβ6 is an RGD-binding and uPAR-binding receptor. The server produces a cluster dendogram and a heat map using R software for statistical computing and analytical grouping (shown in Fig. 3). Fig. 3a clearly shows the clustering of our integrin αvβ6 model with other RGD-binding and uPAR-binding receptors. The integrins αXβ2 (PDB: 3K6S) and α4β7 (PDB: 3V4V) are known to be leukocyte-specific (and non RGD-binding) receptors, whereas the integrins αvβ3 (PDB: 3IJE), αIIbβ3 (PDB: 3FCS) and α5β1 (PDB: 3VI3) are RGD-binding receptors. Our electrostatic potential similarity analysis suggests common surface electrostatics at work amongst the RGD-binding integrins.

While uPAR is known to functionally interact with β1, β2 and β3 integrin families (Ossowski and Aguirre-Ghiso, 2000), there is no evidence yet to prove that the β7 integrin family binds to uPAR. From Fig. 3b, it is evident that the electrostatic surface potentials among other known uPAR-binding αβ integrins (αXβ2 – PDB: 3K6S, αvβ3 – PDB: 3IJE, αIIbβ3 – PDB: 3FCS, α5β1 – PDB: 3VI3)

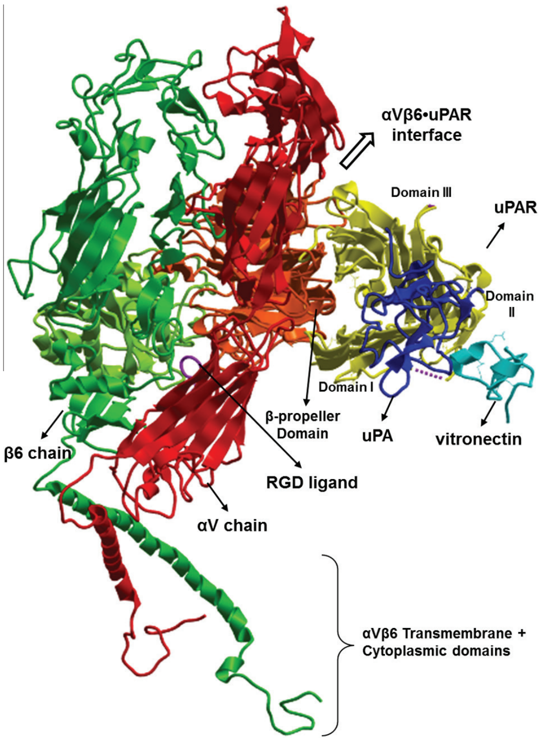


**Fig.4.** Integrin αvβ6·uPAR complex with bound uPA and VN. The docked αvβ6·uPAR complex along with uPA and VN are shown in cartoon representation, with αv chain coloured in red, β6 chain in green, uPAR in yellow, uPA in blue and VN in grey. The αvβ6 domains of the docked complex (shown in a lighter shade) were superposed onto the complete integrin αvβ6 model (along with transmembrane and cytoplasmic domains). The structure of uPA and VN in complex with uPAR was obtained from the PDB (ID: 3BT1) and superimposed onto uPAR in our docked complex. The RGD ligand (in magenta) was obtained from αvβ3 structure (PDB: 1L5G (Xiong et al., 2002)) and superposed onto the integrin model to show the RGD-binding site on integrin. The αvβ6·uPAR docked complex shows that the β-propeller region of α-chain of integrin αvβ6 interacts with the domain III region of uPAR. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

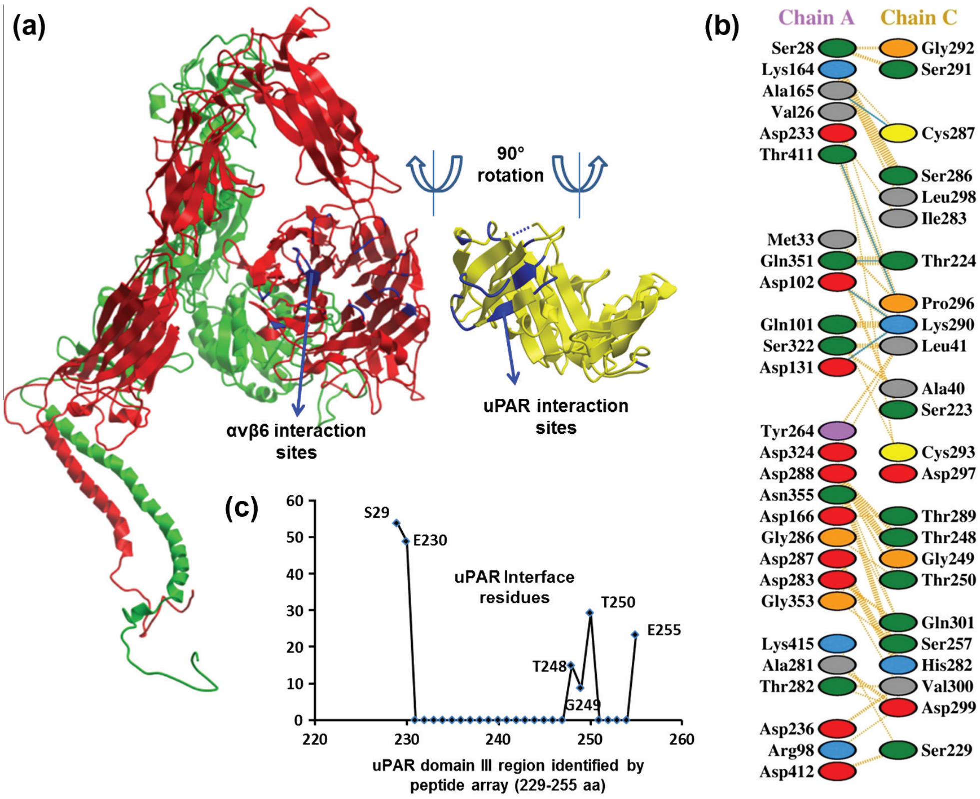


**Fig.5.** The interaction sites of integrin αvβ6-uPAR docked complex. (a) αvβ6 and uPAR rotated by 90° left and right respectively, as indicated. The αvβ6 and uPAR structures are shown in cartoon representation with αv coloured in red, β6 in green and the uPAR in yellow. The αvβ6 and uPAR are rotated by 90° in opposite directions, to show the mutual sites of interaction (coloured in blue) on both structures. (b) Schematic diagram of αv and uPAR interface. The PDBsum interaction analysis represents the interaction residues on either chain with residues shown in different colours based on their properties and the coloured lines joining these residues representing the type of interaction between these residues. (c) A graphical representation of the uPAR binding residues on integrin αvβ6–uPAR interface obtained from docking simulation as a function of residue number using ΔASA. The residue numbers of the uPAR region identified to bind to αvβ6 integrin by PLA and peptide array experiments are plotted on the *x*-axis while ΔASA is on the *y*-axis. The uPAR interface residues and their numbers obtained from the ΔASA analysis of αvβ6–uPAR docked complex are shown. Six out of the 27 residues (identified by PLA and peptide array experiment) are consistent with docking result.

and integrin αvβ6 model developed in this study, are similar. Conversely, Fig. 3b shows that α4β7 (PDB: 3V4V; not known to bind uPAR) does not cluster with other known uPAR-binding αβ integrins and is electrostatically furthest from the integrin αvβ6 model, thereby acting as a control for this analysis. Thus, our MSEP similarity analysis provides substantial evidence that integrin αvβ6 is an uPAR-binding receptor. However, it is not clear whether there is any overlap between the RGD- and uPAR-binding domains or if they are distinct.

### *3.4. αv subunit of αvβ6 integrin interacts with outer surface of uPAR domain III*

The ICM docking simulations of integrin αvβ6 model with the crystal structure of uPAR generated multiple conformations of the docked complexes. Three docking projects with different starting geometries were carried out to ensure that the starting orientation did not bias the final orientation of the binding complex. The final conformations were identified following refinement to minimise the energy and the RMSD of the docked complexes. All three docking experiments result in similar sites of interaction for integrin αvβ6 model and uPAR (data not shown), with the least energy conformation selected for further analysis. These outcomes reaffirm that the selection of different epitopes or potential binding sites of interest on the outer surface of uPAR does not bias the final binding conformation of the docked complex. Experimental reports have shown that integrin α5β1 interacts with domain III region of uPAR (Chaurasia et al., 2006). The surface loop of the β-propeller domain of integrin α-chain in α3β1 and αMβ2 heterodimers were previously reported to functionally associate with uPAR (Simon et al., 2000; Zhang et al., 2003). The refined complex with minimum energy and least RMSD (of all the three docking projects) also reveals similar binding regions wherein the β-propeller region of αv interacts with the outer surface of uPAR domain III. Our docking simulation of integrin αvβ3 (PDB: 3IJE) and uPAR (PDB: 1YWH) also revealed similar sites of interaction, wherein the β-propeller

**Table 3**
Amino acid residue composition for the solvent accessible surface of uPAR.

| aa residue type | | | Surface (% aa composition) |
|---|---|---|---|
| *Polar* | Neutral | | 34.50 |
| | Charged | +ve | 10.70 |
| | | −ve | 13.10 |
| | Total | | 58.30 |
| Non-polar | | | 41.70 |

region of αv from αvβ3 interacts with domain II and domain III of uPAR as shown in Supplementary Fig. S3 (Degryse et al., 2005).

Fig. 4 depicts the results of the *in silico* αvβ6·uPAR docked complex along with the uPA and vitronectin (VN) bound to the uPAR structure, which could represent the binding/complexation model of these proteins *in vivo*. The αvβ6 domain regions (considered for our αvβ6·uPAR docking project) were superposed onto the complete integrin αvβ6 model to visualise the αvβ6·uPAR binding mode of interaction *in vivo*. The crystal structure of the uPAR, uPA and VN complex was obtained from the PDB (ID: 3BT1(Huai et al., 2008)) and superposed onto the position of uPAR in our αvβ6·uPAR docked complex, in order to represent multiple yet ligand-specific binding sites on the uPAR surface. The viability of the proposed αvβ6·uPAR interface, in the presence of uPAR-bound uPA and VN, is thus verified. We note that the RGD-binding site is spatially separated from the uPAR-binding site defined by docking, so that αvβ6 could independently bind fibronection, osteopontin and other ligands at the RGD-binding site.

Fig. 5 represents the sites of interaction on the uPAR and α-chain of the integrin αvβ6 model. The αvβ6 and uPAR structures in the docked complex are shown separated and rotated by 90° (as if opening out the pages of a book) to visualise the mutually interacting sites on both proteins. The PDBsum interaction analysis (Laskowski, 2009) of the docked complex depicts the predicted sites of interaction with residue numbers coloured based on their aminoacid properties (positively charged: H, K, R in blue; negatively charged: D, E in red; neutral: S, T, N, Q in green; aliphatic: A, V, L, I, M in grey; aromatic: F, Y, W in purple; proline and glycine: P, G in orange; cysteine: C in yellow) and bonds between these residues represented as coloured lines (red: salt bridge; yellow: disulphide bonds; blue: hydrogen bonds; orange: non-bonded contacts) for chain A (αv chain) and chain C (uPAR). The docked complex demonstrates a binding mode between the integrin αvβ6 and uPAR, which could be a potential binding conformation *in vivo* for the two proteins to interact and thereby perform biological functions.

Interaction of αvβ6·uPAR was shown by previous study using immunoprecipitation followed by mass spectrometry protein identification (Saldanha et al., 2007). Recent data from our laboratory using proximity ligation assays and peptide array experiments revealed that domains II and III of uPAR interact with αvβ6 (unpublished data). The peptide array has shown that the domain II region (peptides: L172–F189 and C193–E207) and the domain III region of uPAR ranging from S299–N255 (27 residues) interact with αvβ6 integrin. Hence, we performed ASA analysis to obtain buried and surface residues of uPAR in order to eliminate surface inaccessible residues from these peptide segments. The aa residue composition of uPAR shows that the 91% of uPAR residues are surface exposed and that the accessible surface is highly polar with predominantly polar interactions (Table 3).

Docking data reduced the six possible peptide segments known to interact with αvβ6 integrin to the only peptide segment on domain III (27 residues: S299–N255). Of these 27 residues, 5 (C237, M241, C244, L245, and A247) in the identified domain III region are found to be buried in the uPAR structure, with 22 surface residues. Fig. 5c shows that 6 out of 22 surface residues (27.2%) found in this segment are consistent with the docking data (S229, E230, T248, G249, T250, and E255). Integrin α5β1 is known to interact with the domain III region of uPAR via the peptide GCATASMCQ (referred to as "240–248") using co-immunoprecipitation experiments (Chaurasia et al., 2006). This peptide corresponds to 262–270 in UniProt numbering, used in this study. Our docking results show that two residues on the domain III region of uPAR (C247 and Q248 out of the 9 residues known to bind integrin α5β1) interact with integrin αv residues (G234 and K258) identified by alignment of the homologous interaction sites on integrin α3, α5 and αM chains (Chaurasia et al., 2006; Simon et al., 2000; Zhang et al., 2003). Therefore, the predicted binding modes of interaction of our integrin αvβ6·uPAR docked complex is in accordance with experimental data.

## 4. Conclusions

Integrin αvβ6, a transmembrane protein binds fibronectin, osteopontin, and LAP of TGF-β to perform various essential biological functions. The specific interaction between integrin αvβ6 and uPAR is known to be a key step in cancer progression. An investigation on the distinctive yet descriptive structural properties of integrin αvβ6 in comparison with other X-ray crystal structures of integrins known to interact with uPAR, gives a better understanding of integrin αvβ6 mediated regulation of the plasminogen activator system. The consensus of knowledge gleaned between structural analysis and docking information suggests the mode of interaction of uPAR to integrin αvβ6. Although, functional attribution of structural interface between two proteins is often a non-trivial task, we believe that the convergence of our structural analysis and docking data with previous experimental results aids in understanding the molecular basis of αvβ6·uPAR interactions. Six (S229, E230, T248, G249, T250, and E255) out of 27 aa in the identified uPAR domain III binding region is consistent with docking data. Therefore, the αvβ6 structural model and αvβ6-uPAR molecular docking simulations are informative in eliminating possible false positives obtained from experimental data and thereby identify the high-affinity potential site of interaction between αvβ6 and uPAR in 3D space. These observations have implications for the abrogation of the αvβ6-uPAR interaction as a potential therapeutic target in cancer management, using specific chemical inhibitors as demonstrated for α5β1 (Chaurasia et al., 2009).

## Acknowledgments

G.S. is the recipient of an iMQRES PhD scholarship from Macquarie University. We are grateful to Dr. Michael Agrez, John Hunter Hospital, NSW, for his valuable suggestions.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2014.01.001.

## References

Abagyan, R.A.T., Kuznetsov, M.M.D.A., 1994. ICM: a new method for protein modelling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. 15, 488–506.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Arnaout, M.A., Goodman, S.L., Xiong, J.P., 2007. Structure and mechanics of integrin-based cell adhesion. Curr. Opin. Cell Biol. 19, 495–507.

Bandyopadhyay, A., Raghavan, S., 2009. Defining the role of integrin alphavbeta6 in cancer. Curr. Drug Targets 10, 645–652.

Barczyk, M., Carracedo, S., Gullberg, D., 2010. Integrins. Cell Tissue Res. 339, 269–280.

Bates, R.C., Bellovin, D.I., Brown, C., Maynard, E., Wu, B., Kawakatsu, H., Sheppard, D., Oettgen, P., Mercurio, A.M., 2005. Transcriptional activation of integrin beta6 during the epithelial–mesenchymal transition defines a novel prognostic indicator of aggressive colon carcinoma. J. Clin. Invest. 115, 339–347.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bhattacharya, A., Tejero, R., Montelione, G.T., 2007. Evaluating protein structures determined by structural genomics consortia. Proteins 66, 778–795.

Binder, M., Trepel, M., 2009. Drugs targeting integrins for cancer therapy. Expert Opin. Drug Dis. 4, 229–241.

Blomberg, N., Gabdoulline, R.R., Nilges, M., Wade, R.C., 1999. Classification of protein sequences by homology modelling and quantitative analysis of electrostatic similarity. Proteins 37, 379–387.

Breuss, J.M., Gallo, J., DeLisser, H.M., Klimanskaya, I.V., Folkesson, H.G., Pittet, J.F., Nishimura, S.L., Aldape, K., Landers, D.V., Carpenter, W., et al., 1995. Expression of the beta 6 integrin subunit in development, neoplasia and tissue repair suggests a role in epithelial remodelling. J. Cell Sci. 108 (Pt 6), 2241–2251.

Busk, M., Pytela, R., Sheppard, D., 1992. Characterization of the integrin alpha v beta 6 as a fibronectin-binding protein. J. Biol. Chem. 267, 5790–5796.

Campbell, I.D., Humphries, M.J., 2011. Integrin structure, activation, and interactions. Cold Spring Harb. Perspect. Biol. 3.

Chaurasia, P., Aguirre-Ghiso, J.A., Liang, O.D., Gardsvoll, H., Ploug, M., Ossowski, L., 2006. A region in urokinase plasminogen receptor domain III controlling a functional association with alpha5beta1 integrin and tumor growth. J. Biol. Chem. 281, 14852–14863.

Chaurasia, P., Mezei, M., Zhou, M.M., Ossowski, L., 2009. Computer aided identification of small molecules disrupting uPAR/alpha5beta1–integrin interaction: a new paradigm for metastasis prevention. PLoS One 4, e4617.

Degryse, B., Resnati, M., Czekay, R.P., Loskutoff, D.J., Blasi, F., 2005. Domain 2 of the urokinase receptor contains an integrin-interacting epitope with intrinsic signaling activity: generation of a new integrin inhibitor. J. Biol. Chem. 280, 24792–24803.

Diago, L.A., Morell, P., Aguilera, L., Moreno, E., 2007. Setting up a large set of protein–ligand PDB complexes for the development and validation of knowledge-based docking algorithms. BMC Bioinformatics 8, 310.

Eden, G., Archinti, M., Furlan, F., Murphy, R., Degryse, B., 2011. The urokinase receptor interactome. Curr. Pharm. Des. 17, 1874–1889.

Giancotti, F.G., Ruoslahti, E., 1999. Integrin signaling. Science 285, 1028–1032.

Hooft, R.W., Vriend, G., Sander, C., Abola, E.E., 1996. Errors in protein structures. Nature 381, 272.

Huai, Q., Zhou, A., Lin, L., Mazar, A.P., Parry, G.C., Callahan, J., Shaw, D.E., Furie, B., Furie, B.C., Huang, M., 2008. Crystal structures of two human vitronectin, urokinase and urokinase receptor complexes. Nat. Struct. Mol. Biol. 15, 422–423.

Humphries, M.J., 2000. Integrin structure. Biochem. Soc. Trans. 28, 311–339.

Huynh, T., Khan, J.M., Ranganathan, S., 2011. A comparative structural bioinformatics analysis of inherited mutations in beta-D-Mannosidase across multiple species reveals a genotype–phenotype correlation. BMC Genomics 12 (Suppl. 3), S22.

Hynes, R.O., 2002. Integrins: bidirectional, allosteric signaling machines. Cell 110, 673–687.

Hynes, R.O., 2004. The emergence of integrins: a personal and historical perspective. Matrix Biol. 23, 333–340.

Khan, J.M., Ranganathan, S., 2009. A multi-species comparative structural bioinformatics analysis of inherited mutations in alpha-D-mannosidase reveals strong genotype–phenotype correlation. BMC Genomics 10 (Suppl. 3), S33.

Kim, C., Ye, F., Ginsberg, M.H., 2011. Regulation of integrin activation. Annu. Rev. Cell Dev. Biol. 27, 321–345.

Laskowski, R.A., 2009. PDBsum new things. Nucleic Acids Res. 37, D355–359.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., Thornton, J.M., 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR 8, 477–486.

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379–400.

Llinas, P., Le Du, M.H., Gardsvoll, H., Dano, K., Ploug, M., Gilquin, B., Stura, E.A., Menez, A., 2005. Crystal structure of the human urokinase plasminogen activator receptor bound to an antagonist peptide. EMBO J. 24, 1655–1663.

Lo Conte, L., Chothia, C., Janin, J., 1999. The atomic structure of protein–protein recognition sites. J. Mol. Biol. 285, 2177–2198.

Lu, X., Lu, D., Scully, M., Kakkar, V., 2008. The role of integrins in cancer and the development of anti-integrin therapeutic agents for cancer therapy. Perspect. Med. Chem. 2, 57–73.

Margadant, C., Sonnenberg, A., 2010. Integrin-TGF-beta crosstalk in fibrosis, cancer and wound healing. EMBO Rep. 11, 97–105.

Morgan, M.R., Thomas, G.J., Russell, A., Hart, I.R., Marshall, J.F., 2004. The integrin cytoplasmic-tail motif EKQKVDLSTDC is sufficient to promote tumor cell invasion mediated by matrix metalloproteinase (MMP)-2 or MMP-9. J. Biol. Chem. 279, 26533–26539.

Morris, D.G., Huang, X., Kaminski, N., Wang, Y., Shapiro, S.D., Dolganov, G., Glick, A., Sheppard, D., 2003. Loss of integrin alpha(v)beta6-mediated TGF-beta activation causes Mmp12-dependent emphysema. Nature 422, 169–173.

Munger, J.S., Huang, X., Kawakatsu, H., Griffiths, M.J., Dalton, S.L., Wu, J., Pittet, J.F., Kaminski, N., Garat, C., Matthay, M.A., Rifkin, D.B., Sheppard, D., 1999. The integrin alpha v beta 6 binds and activates latent TGF beta 1: a mechanism for regulating pulmonary inflammation and fibrosis. Cell 96, 319–328.

Nagae, M., Re, S., Mihara, E., Nogi, T., Sugita, Y., Takagi, J., 2012. Crystal structure of alpha5beta1 integrin ectodomain: atomic details of the fibronectin receptor. J. Cell Biol. 197, 131–140.

Neves, M.A., Totrov, M., Abagyan, R., 2012. Docking and scoring with ICM: the benchmarking results and strategies for improvement. J. Comput. Aided Mol. Des. 26, 675–686.

Ossowski, L., Aguirre-Ghiso, J.A., 2000. Urokinase receptor and integrin partnership: coordination of signaling for cell adhesion, migration and growth. Curr. Opin. Cell Biol. 12, 613–620.

Porollo, A., Meller, J., 2007. Prediction-based fingerprints of protein–protein interactions. Proteins 66, 630–645.

Ranganathan, S., 2001. Molecular modelling on the web. Biotechniques 30, 50–52.

Richter, S., Wenzel, A., Stein, M., Gabdoulline, R.R., Wade, R.C., 2008. WebPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acids Res. 36, W276–280.

Saldanha, R.G., Molloy, M.P., Bdeir, K., Cines, D.B., Song, X., Uitto, P.M., Weinreb, P.H., Violette, S.M., Baker, M.S., 2007. Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. J. Proteome Res. 6, 1016–1028.

Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815.

Sanchez, R., Sali, A., 1997. Advances in comparative protein–structure modelling. Curr. Opin. Struct. Biol. 7, 206–214.

Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9, 56–68.

Simon, D.I., Wei, Y., Zhang, L., Rao, N.K., Xu, H., Chen, Z., Liu, Q., Rosenberg, S., Chapman, H.A., 2000. Identification of a urokinase receptor–integrin interaction site. Promiscuous regulator of integrin function. J. Biol. Chem. 275, 10228–10234.

Smith, H.W., Marshall, C.J., 2010. Regulation of cell signalling by uPAR. Nat. Rev. Mol. Cell Biol. 11, 23–36.

Sowmya, G., Anita, S., Kangueane, P., 2011. Insights from the structural analysis of protein heterodimer interfaces. Bioinformation 6, 137–143.

Takada, Y., Ye, X., Simon, S., 2007. The integrins. Genome Biol. 8, 215.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Tng, E., Tan, S.M., Ranganathan, S., Cheng, M., Law, S.K., 2004. The integrin alpha L beta 2 hybrid domain serves as a link for the propagation of activation signal from its stalk regions to the I-like domain. J. Biol. Chem. 279, 54334–54339.

Tsodikov, O.V., Record Jr., M.T., Sergeev, Y.V., 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J. Comput. Chem. 23, 600–609.

UniProt Consortium, 2014. Activities at the Universal Protein Resource (UniProt). Nucl. Acids Res. 42, D191–D198.

Van Aarsen, L.A., Leone, D.R., Ho, S., Dolinski, B.M., McCoon, P.E., LePage, D.J., Kelly, R., Heaney, G., Rayhorn, P., Reid, C., Simon, K.J., Horan, G.S., Tao, N., Gardner, H.A., Skelly, M.M., Gown, A.M., Thomas, G.J., Weinreb, P.H., Fawell, S.E., Violette, S.M., 2008. Antibody-mediated blockade of integrin alpha v beta 6 inhibits tumor progression *in vivo* by a transforming growth factor-beta-regulated mechanism. Cancer Res. 68, 561–570.

Xie, C., Zhu, J., Chen, X., Mi, L., Nishida, N., Springer, T.A., 2010. Structure of an integrin with an alphaI domain, complement receptor type 4. EMBO J. 29, 666–679.

Xiong, J.P., Stehle, T., Diefenbach, B., Zhang, R., Dunker, R., Scott, D.L., Joachimiak, A., Goodman, S.L., Arnaout, M.A., 2001. Crystal structure of the extracellular segment of integrin alpha Vbeta3. Science 294, 339–345.

Xiong, J.P., Stehle, T., Zhang, R., Joachimiak, A., Frech, M., Goodman, S.L., Arnaout, M.A., 2002. Crystal structure of the extracellular segment of integrin alpha Vbeta3 in complex with an Arg-Gly-Asp ligand. Science 296, 151–155.

Xiong, J.P., Mahalingham, B., Alonso, J.L., Borrelli, L.A., Rui, X., Anand, S., Hyman, B.T., Rysiok, T., Muller-Pompalla, D., Goodman, S.L., Arnaout, M.A., 2009. Crystal structure of the complete integrin alphaVbeta3 ectodomain plus an alpha/beta transmembrane fragment. J. Cell Biol. 186, 589–600.

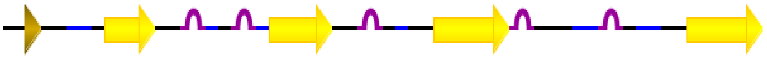
Yang, J., Ma, Y.Q., Page, R.C., Misra, S., Plow, E.F., Qin, J., 2009. Structure of an integrin alphaIIb beta3 transmembrane-cytoplasmic heterocomplex provides insight into integrin activation. Proc. Natl. Acad. Sci. USA 106, 17729–17734.

Yu, Y., Zhu, J., Mi, L.Z., Walz, T., Sun, H., Chen, J., Springer, T.A., 2012. Structural specializations of alpha(4)beta(7), an integrin that mediates rolling adhesion. J. Cell Biol. 196, 131–146.

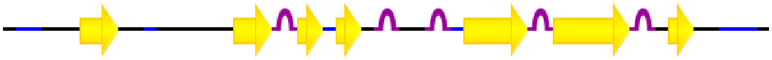
Zhang, F., Tom, C.C., Kugler, M.C., Ching, T.T., Kreidberg, J.A., Wei, Y., Chapman, H.A., 2003. Distinct ligand binding sites in integrin alpha3beta1 regulate matrix adhesion and cell–cell contact. J. Cell Biol. 163, 177–188.

Zhu, J., Luo, B.H., Xiao, T., Zhang, C., Nishida, N., Springer, T.A., 2008. Structure of a complete integrin ectodomain in a physiologic resting state and activation and deactivation by applied forces. Mol. Cell 32, 849–861.

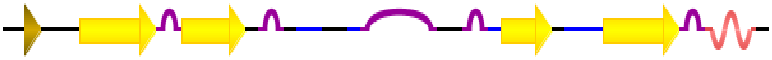
## Supplementary data files



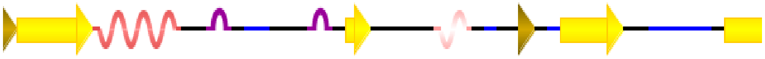
```
sp|P06756|ITAV_HUMAN      FNLDVDSPAEYSGPEGSYFGFAVDFFVPSASSRMFLLVGAPKANTTQPGIVEGGQVLKCD
3IJE_A                    FNLDVDSPAEYSGPEGSYFGFAVDFFVPSASSRMFLLVGAPKANTTQPGIVEGGQVLKCD
                          ************************************************************
```



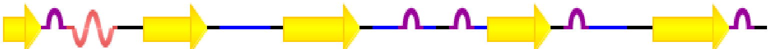
```
sp|P06756|ITAV_HUMAN      WSSTRRCQPIEFDATGNRDYAKDDPLEFKSHQWFGASVRSKQDKILACAPLYHWRTEMKQ
3IJE_A                    WSSTRRCQPIEFDATGNRDYAKDDPLEFKSHQWFGASVRSKQDKILACAPLYHWRTEMKQ
                          ************************************************************
```



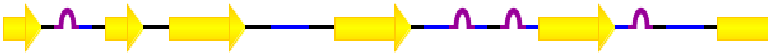
```
sp|P06756|ITAV_HUMAN      EREPVGTCFLQDGTKTVEYAPCRSQDIDADGQGFCQGGFSIDFTKADRVLLGGPGSFYWQ
3IJE_A                    EREPVGTCFLQDGTKTVEYAPCRSQDIDADGQGFCQGGFSIDFTKADRVLLGGPGSFYWQ
                          ************************************************************
```



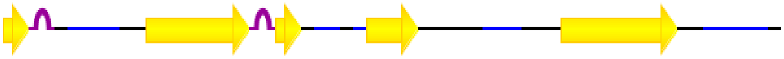
```
sp|P06756|ITAV_HUMAN      GQLISDQVAEIVSKYDPNVYSIKYNNQLATRTAQAIFDDSYLGYSVAVGDFNGDGIDDFV
3IJE_A                    GQLISDQVAEIVSKYDPNVYSIKYNNQLATRTAQAIFDDSYLGYSVAVGDFNGDGIDDFV
                          ************************************************************
```



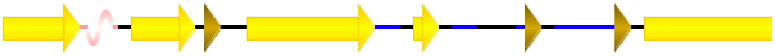
```
sp|P06756|ITAV_HUMAN      SGVPRAARTLGMVYIYDGKNMSSLYNFTGEQMAAYFGFSVAATDINGDDYADVFIGAPLF
3IJE_A                    SGVPRAARTLGMVYIYDGKNMSSLYNFTGEQMAAYFGFSVAATDINGDDYADVFIGAPLF
                          ************************************************************
```



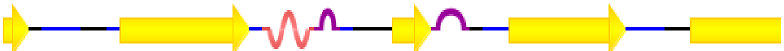
```
sp|P06756|ITAV_HUMAN      MDRGSDGKLQEVGQVSVSLQRASGDFQTTKLNGFEVFARFGSAIAPLGDLDQDGFNDIAI
3IJE_A                    MDRGSDGKLQEVGQVSVSLQRASGDFQTTKLNGFEVFARFGSAIAPLGDLDQDGFNDIAI
                          ************************************************************
```



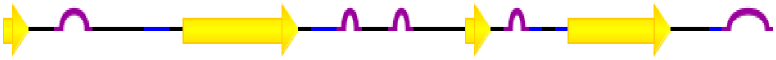
```
sp|P06756|ITAV_HUMAN      AAPYGGEDKKGIVYIFNGRSTGLNAVPSQILEGQWAARSMPPSFGYSMKGATDIDKNGYP
3IJE_A                    AAPYGGEDKKGIVYIFNGRSTGLNAVPSQILEGQWAARSMPPSFGYSMKGATDIDKNGYP
                          ************************************************************
```



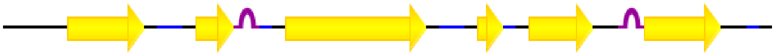
```
sp|P06756|ITAV_HUMAN      DLIVGAFGVDRAILYRARPVITVNAGLEVYPSILNQDNKTCSLPGTALKVSCFNVRFCLK
3IJE_A                    DLIVGAFGVDRAILYRARPVITVNAGLEVYPSILNQDNKTCSLPGTALKVSCFNVRFCLK
                          ************************************************************
```



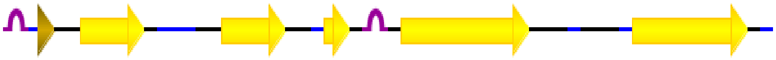
```
sp|P06756|ITAV_HUMAN      ADGKGVLPRKLNFQVELLLDKLKQKGAIRRALFLYSRSPSHSKNMTISRGGLMQCEELIA
3IJE_A                    ADGKGVLPRKLNFQVELLLDKLKQKGAIRRALFLYSRSPSHSKNMTISRGGLMQCEELIA
                          ************************************************************
```



```
sp|P06756|ITAV_HUMAN      YLRDESEFRDKLTPITIFMEYRLDYRTAADTTGLQPILNQFTPANISRQAHILLDCGEDN
3IJE_A                    YLRDESEFRDKLTPITIFMEYRLDYRTAADTTGLQPILNQFTPANISRQAHILLDCGEDN
                          ************************************************************
```

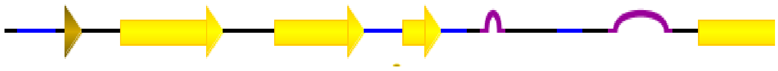


```
sp|P06756|ITAV_HUMAN      VCKPKLEVSVDSDQKKIYIGDDNPLTLIVKAQNQGEGAYEAELIVSIPLQADFIGVVRNN
3IJE_A                    VCKPKLEVSVDSDQKKIYIGDDNPLTLIVKAQNQGEGAYEAELIVSIPLQADFIGVVRNN
                          ************************************************************
```

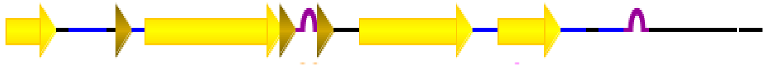


```
sp|P06756|ITAV_HUMAN      EALARLSCAFKTENQTRQVVCDLGNPMKAGTQLLAGLRFSVHQQSEMDTSVKFDLQIQSS
3IJE_A                    EALARLSCAFKTENQTRQVVCDLGNPMKAGTQLLAGLRFSVHQQSEMDTSVKFDLQIQSS
                          ************************************************************
```

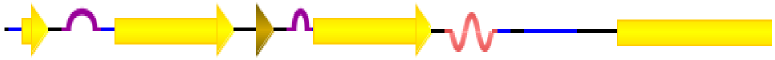
## Supplementary data



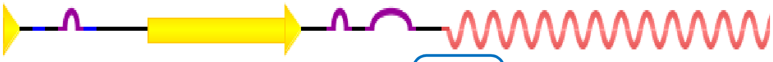
```
sp|P06756|ITAV_HUMAN      NLFDKVSPVVSHKVDLAVLAAVEIRGVSSPDHVFLPIPNWEHKENPETEEDVGPVVQHIY
3IJE_A                    NLFDKVSPVVSHKVDLAVLAAVEIRGVSSPDHVFLPIPNWEHKENPETEEDVGPVVQHIY
                          ************************************************************
```



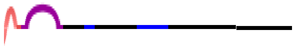
```
sp|P06756|ITAV_HUMAN      ELRNNGPSSFSKAMLHLQWPYKYNNNTLLYILHYDIDGPMNCTSDMEINPLRIKISSLDI
3IJE_A                    ELRNNGPSSFSKAMLHLQWPYKYNNNTLLYILHYDIDGPMNCTSDMEINPLRIKISSLDI
                          ************************************************************
```



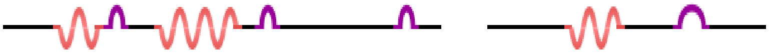
```
sp|P06756|ITAV_HUMAN      HTLGCGVAQCLKIVCQVGRLDRGKSAILYVKSLLWTETFMNKENQNHSYSLKSSASFNVI
3IJE_A                    HTLGCGVAQCLKIVCQVGRLDRGKSAILYVKSLLWTETFMNKENQNHSYSLKSSASFNVI
                          ************************************************************
```



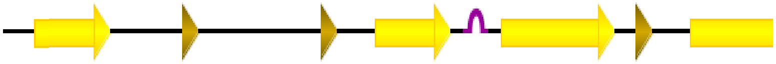
```
sp|P06756|ITAV_HUMAN      EFPYKNLPIEDITNSTLVTTNVTWGIQPAPMPVPVWVIILAVLAGLLLLAVLVFVMYRMG
3IJE_A                    EFPYKNLPIEDITNSTLVTTNVTWGIQPAPMPVPVWVIG
2KNC_A                                           GAMGSEERAIPIWWVLVGVLGGLLLLTILVLAMWKVG
                          ********************************:*:* :::.**.*****::**:.*:::*
```



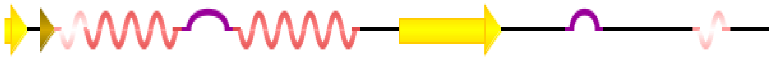
```
sp|P06756|ITAV_HUMAN      FFKRVRPPQEEQEREQLQPHENGEGNSET
2KNC_A                    -FKRNRPPLEEDD-------EEG
                           *** *** **::       *:*
```



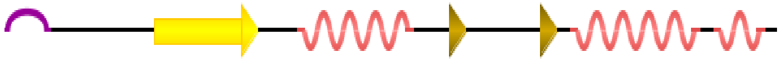
```
sp|P18564|ITB6_HUMAN      ---GCALGGAETCEDCLLIGPQCAWCAQENFTHPSGVGERCDTPANLLAKGCQLNFIENP
3IJE_B                    GPNICTTRGVSSCQQCLAVSPMCAWCSDE---ALPlgsPRCDLKENLLKDNCAPESIEFP
                              *:  *..:*::** :.* ****::*          ***   *** ..*  : ** *
```



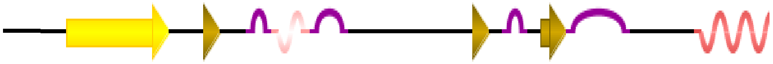
```
sp|P18564|ITB6_HUMAN      VSQVEILKNKPLSVGRQKNSSDIVQIAPQSLILKLRPGGAQTLQVHVRQTEDYPVDLYYL
3IJE_B                    VSEARVLEDRPLSDKGSGDSSQVTQVSPQRIALRLRPDDSKNFSIQVRQVEDYPVDIYYL
                          **:..:*:::***   . :**::.*::** : *:***..::.:.::***.******:***
```



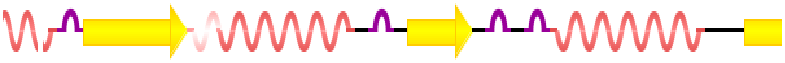
```
sp|P18564|ITB6_HUMAN      MDLSASMDDDLNTIKELGSRLSKEMSKLTSNFRLGFGSFVEKPVSPFVKTTPE-EIANPC
3IJE_B                    MDLSYSMKDDLWSIQNLGTKLATQMRKLTSNLRIGFGAFVDKPVSPYMYISPPEALENPC
                          **** **.*** :*::**::*:.:* *****:*:***:**:*****::  :*   : ***
```



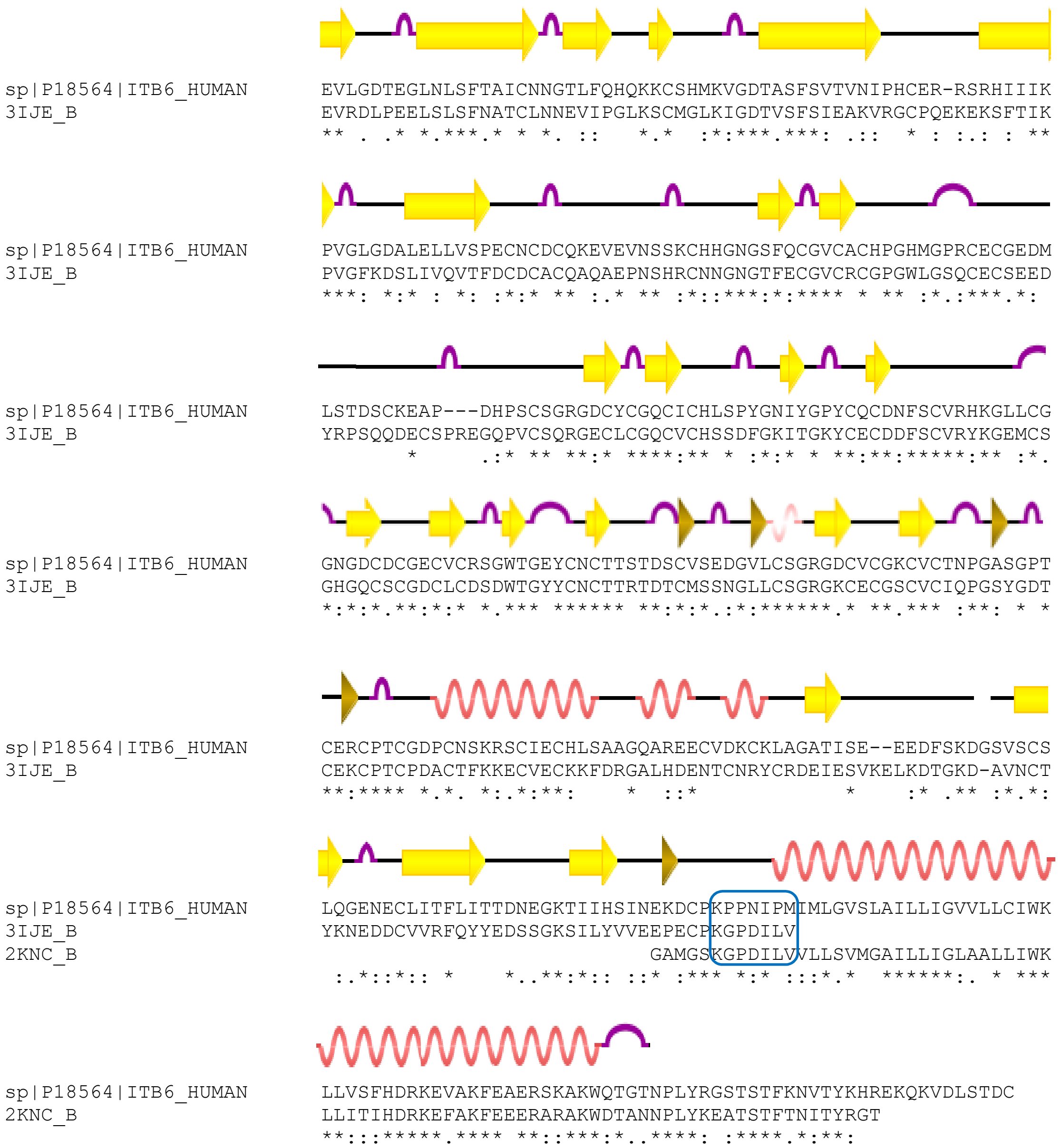
```
sp|P18564|ITB6_HUMAN      SSIPYFCLPTFGFKHILPLTNDAERFNEIVKNQKISANIDTPEGGFDAIMQAAVCKEKIG
3IJE_B                    YDMKTTCLPMFGYKHVLTLTDQVTRFNEEVKKQSVSRNRDAPEGGFDAIMQATVCDEKIG
                           .:   *** **:**:*.**::. **** **:*.:* * *:***********:**.****
```



```
sp|P18564|ITB6_HUMAN      WRNDSLHLLVFVSDADSHFGMDSKLAGIVIPNDGLCHLDSKNEYSMSTVLEYPTIGQLID
3IJE_B                    WRNDASHLLVFTTDAKTHIALDGRLAGIVQPNDGQCHVGSDNHYSASTTMDYPSLGLMTE
                          ****: *****.:**.:*:.:*.:***** **** **:.*.*.** **.::**::* : :
```



```
sp|P18564|ITB6_HUMAN      KLVQNNVLLIFAVTQEQVHLYENYAKLIPGATVGLLQKDSGNILQLIISAYEELRSEVEL
3IJE_B                    KLSQKNINLIFAVTENVVNLYQNYSELIPGTTVGVLSMDSSNVLQLIVDAYGKIRSKVEL
                          ** *:*: ******:: *:**:**::****:***:*. **.*:****:.** ::**:***
```

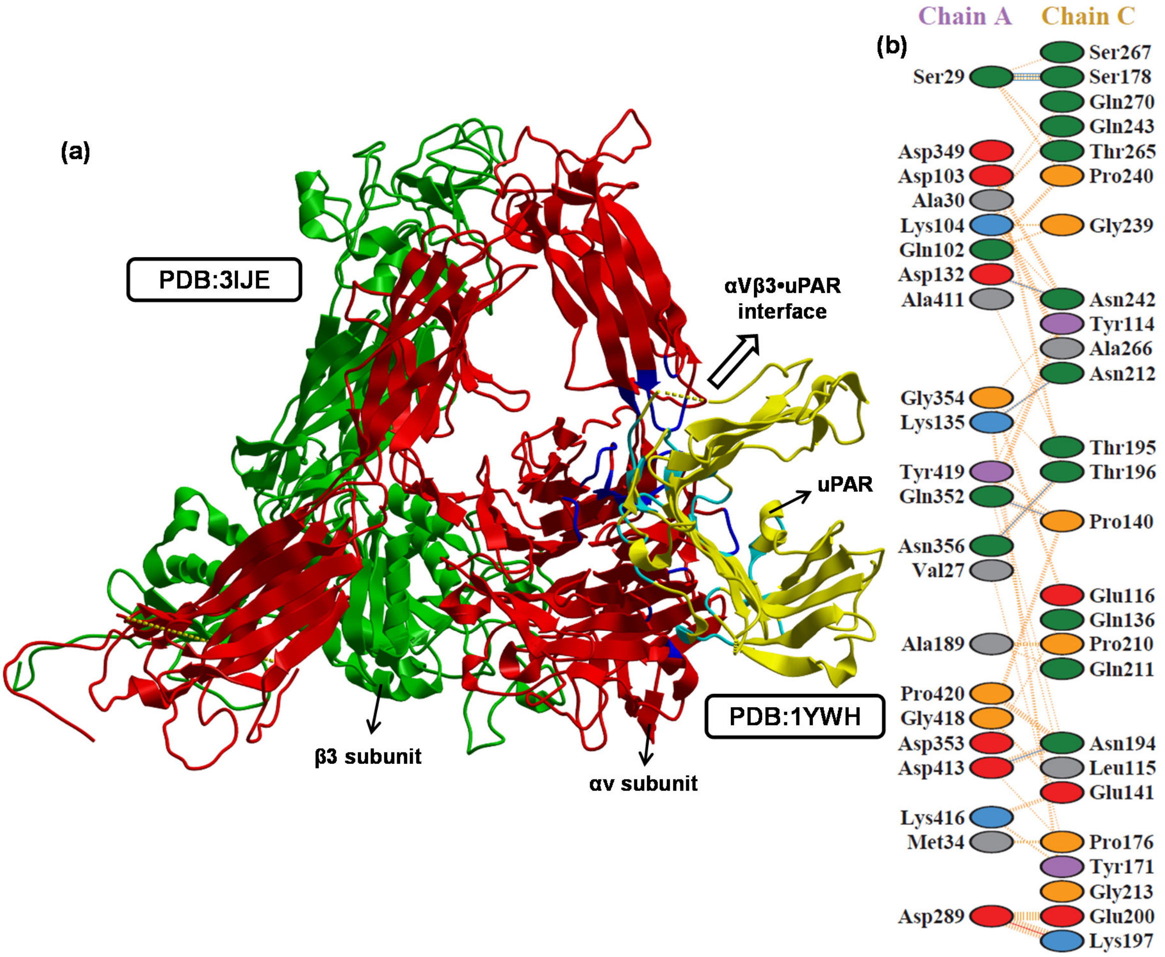


```
sp|P18564|ITB6_HUMAN      EVLGDTEGLNLSFTAICNNGTLFQHQKKCSHMKVGDTASFSVTVNIPHCER-RSRHIIIK
3IJE_B                    EVRDLPEELSLSFNATCLNNEVIPGLKSCMGLKIGDTVSFSIEAKVRGCPQEKEKSFTIK
                          ** . .* *.***.* * *. ::   *.*  :*:***.***: .::  * : :.: : **

sp|P18564|ITB6_HUMAN      PVGLGDALELLVSPECNCDCQKEVEVNSSKCHHGNGSFQCGVCACHPGHMGPRCECGEDM
3IJE_B                    PVGFKDSLIVQVTFDCDCACQAQAEPNSHRCNNGNGTFECGVCRCGPGWLGSQCECSEED
                          ***: *:* : *: :*:* ** :.* ** :*::***:*:**** * ** :*.:***.*:

sp|P18564|ITB6_HUMAN      LSTDSCKEAP---DHPSCSGRGDCYCGQCICHLSPYGNIYGPYCQCDNFSCVRHKGLLCG
3IJE_B                    YRPSQQDECSPREGQPVCSQRGECLCGQCVCHSSDFGKITGKYCECDDFSCVRYKGEMCS
                                 *     .:* ** **:* ****:** * :*:* * **:**:*****:** :*.

sp|P18564|ITB6_HUMAN      GNGDCDCGECVCRSGWTGEYCNCTTSTDSCVSEDGVLCSGRGDCVCGKCVCTNPGASGPT
3IJE_B                    GHGQCSCGDCLCDSDWTGYYCNCTTRTDTCMSSNGLLCSGRGKCECGSCVCIQPGSYGDT
                          *:*:*.**:*:* *.*** ****** **:*:*.:*:******.* **.*** :**: * *

sp|P18564|ITB6_HUMAN      CERCPTCGDPCNSKRSCIECHLSAAGQAREECVDKCKLAGATISE--EEDFSKDGSVSCS
3IJE_B                    CEKCPTCPDACTFKKECVECKKFDRGALHDENTCNRYCRDEIESVKELKDTGKD-AVNCT
                          **:**** *.*. *:.*:**:    *  ::*            *    :* .** :*.*:

sp|P18564|ITB6_HUMAN      LQGENECLITFLITTDNEGKTIIHSINEKDCPKPPNIPMIMLGVSLAILLIGVVLLCIWK
3IJE_B                    YKNEDDCVVRFQYYEDSSGKSILYVVEEPECPKGPDILV
2KNC_B                                                 GAMGSKGPDILVVLLSVMGAILLIGLAALLIWK
                           :.*::*:: *    *..**:*:: ::* :*** *:* :::*.*  ******:. * ***

sp|P18564|ITB6_HUMAN      LLVSFHDRKEVAKFEAERSKAKWQTGTNPLYRGSTSTFKNVTYKHREKQKVDLSTDC
2KNC_B                    LLITIHDRKEFAKFEEERARAKWDTANNPLYKEATSTFTNITYRGT
                          **:::*****.**** **::***:*..****: :****.*:**:
```

**Figure S1: Alignment between integrin αvβ6 and αvβ3 template sequences (PDB: 3IJE and 2KNC) along with secondary structure.** The dark green and light green highlighted regions represent transmembrane and cytoplasmic domains respectively. The blue highlighted region represents the gaps in the integrin αvβ6 and template sequences. The gaps were then adjusted by pinching off or adding appropriate aminoacid residues at the (end) loop regions; of the adjacent sequence such that there is less (or no) hindrance to the secondary structure conformation or other possible interactions of such residues with the core residues. The overlapping residues between inter-domains are shown in blue boxes, with the grey highlighted regions not used in the model building. The red highlighted residues represent the interface residues between the αv and β3 chains (ectodomain) of integrin αvβ3 template structure (PDB: 3IJE). *(The number of interface residues in the 3IJE_B protein subunit is 59, out of which 30 residues (50.85%) are conserved in the human integrin β6 protein, 14 residues (23.73%) are conservatively substituted, 4 residues (6.78%) are semi-conservatively substituted while 11 residues (18.64%) (I167T, Y657L, R216I, Q267I, L294Q, D477M, Y478L, R479S, R563S, R666T, Y594S) are dissimilar. The interface residues do not form hydrogen bonds; hence these non-bonded contacts may be due to hydrophobic interactions or van der Waals forces).*

**Supplementary data file 2:** The coordinates of the final model of integrin αvβ6 heterodimer in PDB format is available at:
http://www.sciencedirect.com/science/article/pii/S1047847714000021#m0025

As this file is very big, it has not been included in the thesis.



**Supplementary Figure S3: The interaction sites of integrin αvβ3•uPAR docked complex. (a)** The αvβ3•uPAR docked complex is shown in cartoon representation with αv coloured in red, β3 in green and uPAR in yellow. The interaction sites on αv subunit of integrin αvβ3 and domain II and domain III of uPAR structures are coloured in dark blue and light blue respectively. **(b) Schematic diagram of αv (chain A) and uPAR (chain C; residues renumbered as per UniProt: Q03405) interface.** The PDBsum interaction analysis represents the interaction residues on either chain with residues shown in different colours based on their properties and the coloured lines joining these residues representing the type of interaction between these residues.

## 5.3 Conclusions

The specific interaction between integrin αvβ6 and uPAR is known to be a key step in cancer progression. A complete 3D structural model of integrin αvβ6 along with extracellular domain, transmembrane domain and cytoplasmic domain has been built. An investigation on the distinctive yet descriptive structural properties of integrin αvβ6 in comparison with other X-ray crystal structures of integrins known to interact with uPAR reveals that αvβ6 integrin is independently an RGD-binding and an uPAR-binding receptor. Docking simulations between the integrin αvβ6 structural model and uPAR protein reveal a single potential interaction site. These observations have implications for the abrogation of the integrin αvβ6•uPAR interaction as a potential therapeutic target in cancer management, using specific chemical inhibitors as demonstrated for α5β1 [325]. These results provide better understanding of integrin αvβ6 mediated regulation of the plasminogen activator system and help gain insights into integrin αvβ6•uPAR interactions.

The consensus of knowledge gleaned between structural analysis and docking information suggests the mode of interaction of uPAR to integrin αvβ6. Although, functional attribution of structural interface between two proteins is often a non-trivial task, the convergence of our structural analysis and docking data with previous experimental results is believed to aid in understanding the molecular basis of PPIs.

This preliminary application of PPIs helped verify the abundance of polar interactions at the integrin αv-β6 subunit interface and also characterised the integrin protein interfaces.

# Chapter 6: Conclusions and future directions

## 6.1 Summary

Protein-protein interactions (PPI) are common in molecular catalysis, regulation, human genetics, human diseases and biotechnology applications. Therefore, it is of interest to study the molecular basis of PPI. Structural understanding of protein complexes using representative sets of known complexes have improved our understanding of this phenomenon over the last five decades. These observations laid the foundation to the formulation of rigid body protein-protein docking and analysis for application in molecular cellular biology. However, there is a huge scope for improvement in docking and post model analysis of protein complexes for mimicking molecular events. Hence, there is a need to revisit this phenomenon using an updated set of known structural complexes.

A review on PPIs focussing on current trends in structural analysis of known proteins to understand their binding is presented (Chapter 1). At the outset, the various experimental techniques used to determine PPIs and their disadvantages have been discussed. The key databases archiving these experimental data, as well as computationally predicted PPI information is documented. The current trends in interaction analyses and prediction, describing the classification of PPIs using various structural and sequence based interface features studied during the past few decades are then reviewed. The variations in PPI datasets created/collected by different groups are then compared and discussed. These datasets that led to the creation of interface databases that provide the data for currently available interaction characterisation and prediction tools/servers are examined. These curated information with online databases and prediction services on PPI have helped to understand its features at large. Nonetheless, this is not yet adequate for a comprehensive understating of the phenomenon.

Extensive studies carried out thus far typically average structural features over diverse datasets. Nevertheless, each PPI complex is specific and selective to binding. Therefore, it is of interest to analyse and study such interface properties in a non-redundant dataset of heterodimer (different subunits) protein-protein complexes (278) available at the PDB (Chapter 3). The relative interface-surface polarity of each complex was estimated in the protein-protein complex dataset to understand the predominant forces driving binding. This showed ~60% of protein complexes as ‘classical’, with surface polarity greater than interface polarity, implying abundant non-polar interactions at the interface (designated as class A). A considerable number of complexes (~40%) have interface polarity greater than surface polarity, implying abundant polar interactions at the interface (designated as class B). Comprehensive analyses of protein complexes show that interface features such as interface area, the relative abundance of polar and non-polar residues, solvent free energy gained upon interface formation, binding energy and percentage of interface charged residues distinguish among class A and class B complexes. It is also subsequently shown that electrostatic visualization maps helps differentiate interface classes among complexes. These novel observations find application in evaluating new interfaces, developing residue-level prediction models, protein-protein docking studies and subunit interface specific inhibitor design as drugs. This could be improved with a collective understanding of molecular functions among protein-protein complexes. Manual curation of the ascribed biological function for each of the 278 protein complexes from known literature showed that all functional categories are represented in the interface classes. The underlying relationship between structures and its features with known biological function is intriguing.

Protein-protein complexes are associated with catalysis, regulation, assembly, immunity and inhibition in a living cell. Therefore, it is of interest to create a comprehensive map through manual grouping between known protein complexes with characterized molecular function in the literature (Chapter 4). The complexes were categorized into major functional groups to identify distinguishing interface features among them. It shows five key features -interface area, interface polar residue abundance (P % - NP %), hydrogen bonds, salt bridges, solvation free energy gain from interface formation, binding energy; that are discriminatory to functional groups. Significant correlations between these interface properties amongst functional groups are also documented. These representative features have implications for the prediction of potential protein function of novel complexes. They find application in understanding diseases through the interpretation of their molecular mechanism.

PPIs underlie the majority of biological processes, signalling, and disease. Small-molecule inhibitors that abolish specific PPIs responsible for diseases have been actively researched and several are currently in clinical use or undergoing clinical trials for gout, dry eye, some cancers, carcinomas and HIV [326]. The relevance of integrin αvβ6•uPAR interaction is known to be crucial for cancer progression. Therefore, it is of interest to analyse and characterize integrin αvβ6 hetero-dimer complex using interface features complemented with known molecular function (Chapter 5). However, the 3D X-ray crystal structure of integrin αvβ6 remains elusive possibly due to its large size, membranous nature and cytoplasmic tail. Nonetheless, the β-propeller region of integrin α-chain is known to interact with uPAR. Recent data from our laboratory using proximity ligation assays (PLA) and peptide arrays showed that domain II and III of uPAR interact with αvβ6 integrin. Therefore, a composite structural model of αvβ6 heterodimer using other known integrin X-ray structures as templates has been built. Subsequently, structural PPI analysis of integrin αvβ6•uPAR interactions was performed using model data with docking simulation for their binding. The interaction region and site on domain III of uPAR and αv subunit is in consensus with experimental data (as detailed in Appendix-1) providing high-affinity potential sites of interaction in 3D space. The molecular basis of integrin αvβ6•uPAR binding using structural data is discussed, for implications as potential therapeutic targets in cancer management.

The work presented in this thesis has utilized various structural analysis tools/software and/or programs to comprehensively investigate binding properties of known protein complexes to gain insights into the molecular principles of PPI. Analysis and grouping of a large number of protein complexes based on interface classes and functional groups lead to the identification of discriminatory features amongst these groups. Incorporation of these combinatorial features is necessary to develop models for protein-protein binding prediction and analysis. Novel observations on modeling membrane protein heterodimer and docking simulation to obtain significant information on key features with integrin αvβ6•uPAR associated cancer progression have been discussed.

## 6.2 Innovations

The thesis highlights original findings and application of protein complexes to study the molecular principles of PPIs and their relation to known biological functions using known structures. Observations on extending the analysis to a larger dataset of protein complexes, for the understanding of the predominant forces driving binding and their relationship to molecular functions, their implication in PPI prediction model and for prediction of protein functions is discussed. PPI analysis on known protein complexes identified discriminatory features amongst interface classes and also among protein functional classes, for applications in the development of PPI binding prediction algorithms, and for the prediction of functions for novel proteins. Novel aspects on the identification of key interactions involved in integrin membrane proteins based on model data, docking simulations and structural analysis of PPI features during cancer progression has been presented as an application in medical sciences.

To the best of my knowledge, the studies described in Chapters 3 and 4: structural analysis of known complexes identified key discriminatory features among interface residue-level classes and reports correlation between PPI structural features and biological function, have not been reported before. Also, the preliminary application of PPI study is the first of its kind, where structural modeling data, docking simulations and PPI analysis of X-ray complex data was used to identify key PPI interactions involved in diseases, such as cancer (Chapter 5).

## 6.3 Significance and contributions

This work emphasises the inherent importance of studying PPIs by the structural analysis of known complexes. A comprehensive overview on PPIs with a focus on interface structural features has been presented in this study. A review of known experimental techniques, mathematically-driven prediction models and information rich curated databases on PPI is reported, in addition to a chronological documentation of several interface features identified by different research groups during the last five decades (Chapter 1).

PPI studies thus far analysed the average hydrophobicity over a diverse set of protein-protein interfaces. This strategy, however, suffers the lack of information for interface stability. Thus, the distribution of hydrophobic features over the individual interfaces also remains unclear. The PPI interface is specific, sensitive and selective for each individual complex. A number of attempts have been made to describe the driving force for PPI using both interface chemical and physical features. However, the compelling reasons for  interactions between heteromeric proteins are not yet evident. Therefore, the description of interfaces using prominent chemical features (such as polarity and hydrophobicity) for each protein-protein complex has direct relevance to their extrapolation in sequences. Hence, known protein-protein complexes from the PDB were classified based on relative interface-surface polarity classes. Key discriminatory features- interface area, interface property abundance (P%-NP%), interface charged residues %, solvent free energy gain upon interface formation

($\Delta^{i}G$), and BE; that are significantly different among these interface classes were identified (Chapter 3).

Protein-protein complexes have critical role in many biological functions. Functional classification of protein-protein complexes into major categories (enzymes, enzyme-inhibitors, regulators, regulator-inhibitors, immune and structural assembly), and PPI analysis, has identified five physicochemical interface features (interface area, interface property abundance (P%-NP%), H-bonds, $\Delta^{i}G$ and BE computed from heterodimer complex structures), that are discriminatory among these functional categories (Chapter 4). PPI features identified and conclusions drawn are discussed to facilitate prediction of novel interaction sites and partners. It also proposes the prediction of biological functions for novel protein-protein complexes.

The inhibition of PPIs for various diseases of therapeutic importance requires the identification of the druggable interface. One such application of PPIs, is the identification of integrin αvβ6•uPAR interface, crucial to cancer progression, as discussed in this thesis (Chapter 5). Our results have implications for drug design to inhibit this specific PPI.

## 6.4 Future directions

The PPI study presented in this thesis could lead to advancements in many directions for the better understanding of molecular principles of PPIs. Structural analysis of protein interface properties to help predict interacting sites and partners is a challenge. Moreover, there are several binding sites in an interacting monomer under *in vivo* conditions. Advances in the analyses of protein interfaces provide insights into the significance of prediction using sequence and structure related information. The classification of complexes based on relative interface-surface polarity has identified key discriminatory features that are significantly different amongst these interface classes (Chapter 3). These observations corroborate the need for classification of protein complexes in determining their combinatorial features and drawing consensus for common patterns in protein-protein recognition. This study should be extended using a combined formulation of residue types and atomic features in future investigation. Furthermore, a detailed analysis of the electrostatic surfaces in each complex, especially in the case of enzymes, could provide explanations for metabolic channelling, as an extension of this work.

The functional classification of complexes and PPI analysis shown in Chapter 4 has identified significantly different PPI features among these categories. Experiments should be formulated to capture these PPI features among functions in future studies. A future extension of this PPI analysis would be to estimate the relative contributions of these PPI features to the function of the protein complex, when the heterodimer dataset is considerably enlarged with new structural information. Relating binding free energy change using combinations of the features derived in the present work will be possible, when a larger dataset is available. The results also have implications for function prediction for orphan proteins, where interacting partners are known and heteromeric complexes can be structurally modelled with high confidence. Nonetheless, the phenomenon and mechanism of gene fusion in multiple domain protein architecture across distant evolutionary history should be considered to the context in future investigations.

Based on the outcomes in Chapter 5, the identified interaction site of integrin αvβ6•uPAR, which is crucial for cancer progression, has implications as potential therapeutic targets in cancer management. The detailed groundwork analysis on integrin αvβ6 structural model and its interactions with other ligands, especially uPAR, can be used to verify these vital interactions using experimental data. In addition, experimental PPI techniques could be used to elucidate these interactions to identify significant pathways affected by these interactions and to ascertain their role. Furthermore, *in vivo* approach of abrogating these interactions in mouse models of colorectal cancer (CRC) can be designed for investigation in future studies.

# References

1. Remy I, Campbell-Valois FX, Michnick SW: **Detection of protein-protein interactions using a simple survival protein-fragment complementation assay based on the enzyme dihydrofolate reductase**. *Nat Protoc* 2007, **2**(9):2120-2125.
2. Huber LA: **Is proteomics heading in the wrong direction?** *Nat Rev Mol Cell Biol* 2003, **4**(1):74-80.
3. Emri T, Tőzsér J: **Protein Biotechnology**. Edited by Tőzsér J: University of Debrecen; 2011.
4. Popiel HA, Burke JR, Strittmatter WJ, Oishi S, Fujii N, Takeuchi T, Toda T, Wada K, Nagai Y: **The aggregation inhibitor peptide qbp1 as a therapeutic molecule for the polyglutamine neurodegenerative diseases**. *J Amino Acids* 2011, **2011**:265084.
5. Sowmya G, Jigisha A, Kangueane P: **Protein-protein complexes**. In: *Protein-protein interactions.* Edited by Kangueane P. New York: Nova Science Publisher; 2010: 1-74.
6. Ozbabacan SE, Engin HB, Gursoy A, Keskin O: **Transient protein-protein interactions**. *Protein Eng Des Sel* 2011, **24**(9):635-648.
7. Sowmya G, Kangueane P: **Protein-protein interactions in hetero-dimer complexes**: Lambert Academic Publishing; 2011.
8. Binder M, Trepel M: **Drugs targeting integrins for cancer therapy**. *Expert Opin Drug Discov* 2009, **4**(3):229-241.
9. Blasi F, Carmeliet P: **uPAR: a versatile signalling orchestrator**. *Nat Rev Mol Cell Biol* 2002, **3**(12):932-943.
10. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain**. *Proc Natl Acad Sci U S A* 1951, **37**(4):205-211.
11. Anfinsen CB: **The formation and stabilization of protein structure**. *Biochem J* 1972, **128**(4):737-749.
12. Lodish H BA, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipurksy SL, Darnell J: **Molecular Cell Biology**, 5 edn. New York: New York: WH Freeman and Company; 2004.
13. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: **Molecular biology of the cell**, 5 edn. New York: Garland Science; 2007.

14. Walsh CT: **Posttranslational modification of proteins: expanding nature's inventory**: Roberts and Co. Publishers; 2006.

15. Nelson DL CM: **Lehninger's Principles of Biochemistry**, 4 edn. New York: New York: W. H. Freeman and Company; 2005.

16. Toyama BH, Hetzer MW: **Protein homeostasis: live long, won't prosper**. *Nat Rev Mol Cell Biol* 2013, **14**(1):55-61.

17. Cambridge SB, Gnad F, Nguyen C, Bermejo JL, Kruger M, Mann M: **Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover**. *J Proteome Res* 2011, **10**(12):5275-5284.

18. Yen HC, Xu Q, Chou DM, Zhao Z, Elledge SJ: **Global protein stability profiling in mammalian cells**. *Science* 2008, **322**(5903):918-923.

19. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G: **The molecular architecture of protein-protein binding sites**. *Curr Opin Struct Biol* 2007, **17**(1):67-76.

20. Robinson CV, Sali A, Baumeister W: **The molecular sociology of the cell**. *Nature* 2007, **450**(7172):973-982.

21. Wells JA, McClendon CL: **Reaching for high-hanging fruit in drug discovery at protein-protein interfaces**. *Nature* 2007, **450**(7172):1001-1009.

22. Janin J, Bahadur RP, Chakrabarti P: **Protein-protein interaction and quaternary structure**. *Q Rev Biophys* 2008, **41**(2):133-180.

23. Jones S, Thornton JM: **Principles of protein-protein interactions**. *Proc Natl Acad Sci U S A* 1996, **93**(1):13-20.

24. Sowmya G, Anita S, Kangueane P: **Insights from the structural analysis of protein heterodimer interfaces**. *Bioinformation* 2011, **6**(4):137-143.

25. Piehler J: **New methodologies for measuring protein interactions in vivo and in vitro**. *Curr Opin Struct Biol* 2005, **15**(1):4-14.

26. Sambourg L, Thierry-Mieg N: **New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size**. *BMC Bioinformatics* 2010, **11**:605.

27. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T *et al*: **A map of the interactome network of the metazoan C. elegans**. *Science* 2004, **303**(5657):540-543.

28. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: **A protein interaction map of Drosophila melanogaster**. *Science* 2003, **302**(5651):1727-1736.

29. Mrowka R, Patzak A, Herzel H: **Is there a bias in proteome research?** *Genome Res* 2001, **11**(12):1971-1973.

30. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML: **Progress and challenges in predicting protein-protein interaction sites**. *Brief Bioinform* 2009, **10**(3):233-246.

31. Chothia C, Janin J: **Principles of protein-protein recognition**. *Nature* 1975, **256**(5520):705-708.

32. Jones S, Thornton JM: **Protein-protein interactions: a review of protein dimer structures**. *Prog Biophys Mol Biol* 1995, **63**(1):31-65.

33. Keskin O, Gursoy A, Ma B, Nussinov R: **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chem Rev* 2008, **108**(4):1225-1244.

34. Chothia C, Wodak S, Janin J: **Role of subunit interfaces in the allosteric mechanism of hemoglobin**. *Proc Natl Acad Sci U S A* 1976, **73**(11):3793-3797.

35. Miller S, Lesk AM, Janin J, Chothia C: **The accessible surface area and stability of oligomeric proteins**. *Nature* 1987, **328**(6133):834-836.

36. Janin J, Chothia C: **The structure of protein-protein recognition sites**. *J Biol Chem* 1990, **265**(27):16027-16030.

37. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches**. *J Mol Biol* 1997, **272**(1):121-132.

38. Xu D, Tsai CJ, Nussinov R: **Hydrogen bonds and salt bridges across protein-protein interfaces**. *Protein Eng* 1997, **10**(9):999-1012.

39. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites**. *J Mol Biol* 1999, **285**(5):2177-2198.

40. Jones S, Marin A, Thornton JM: **Protein domain interfaces: characterization and comparison with oligomeric protein interfaces**. *Protein Eng* 2000, **13**(2):77-82.

41. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **Dissecting subunit interfaces in homodimeric proteins**. *Proteins* 2003, **53**(3):708-719.

42. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J: **A protein domain interaction interface database: InterPare**. *BMC Bioinformatics* 2005, **6**:207.

43. Zhanhua C, Gan JG, Lei L, Sakharkar MK, Kangueane P: **Protein subunit interfaces: heterodimers versus homodimers**. *Bioinformation* 2005, **1**(2):28-39.

44. Pal A, Chakrabarti P, Bahadur R, Rodier F, Janin J: **Peptide segments in protein-protein interfaces**. *J Biosci* 2007, **32**(1):101-111.

45. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N: **PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces**. *Nucleic Acids Res* 2010, **38**(6):e86.

46. Schreiber G, Keating AE: **Protein binding specificity versus promiscuity**. *Curr Opin Struct Biol* 2011, **21**(1):50-61.

47. Morrow JK, Zhang S: **Computational prediction of protein hot spot residues**. *Curr Pharm Des* 2012, **18**(9):1255-1265.

48. Swapna LS, Bhaskara RM, Sharma J, Srinivasan N: **Roles of residues in the interface of transient protein-protein complexes before complexation**. *Sci Rep* 2012, **2**:334.

49. Chen J, Sawyer N, Regan L: **Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area**. *Protein Sci* 2013, **22**(4):510-515.

50. Nooren IM, Thornton JM: **Diversity of protein-protein interactions**. *EMBO J* 2003, **22**(14):3486-3492.

51. Lijnzaad P, Argos P: **Hydrophobic patches on protein subunit interfaces: characteristics and prediction**. *Proteins* 1997, **28**(3):333-343.

52. Smith GR, Sternberg MJ: **Prediction of protein-protein interactions by docking methods**. *Curr Opin Struct Biol* 2002, **12**(1):28-35.

53. Chen H, Zhou HX: **Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data**. *Proteins* 2005, **61**(1):21-35.

54. Szilagyi A, Grimm V, Arakaki AK, Skolnick J: **Prediction of physical protein-protein interactions**. *Phys Biol* 2005, **2**(2):S1-16.

55. Porollo A, Meller J: **Prediction-based fingerprints of protein-protein interactions**. *Proteins* 2007, **66**(3):630-645.

56. Huang B, Schroeder M: **Using protein binding site prediction to improve protein docking**. *Gene* 2008, **422**(1-2):14-21.

57. Park SH, Reyes JA, Gilbert DR, Kim JW, Kim S: **Prediction of protein-protein interaction types using association rule based classification**. *BMC Bioinformatics* 2009, **10**:36.

58. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V: **Predicting protein-protein interface residues using local surface structural similarity**. *BMC Bioinformatics* 2012, **13**:41.

59. Trabuco LG, Lise S, Petsalaki E, Russell RB: **PepSite: prediction of peptide-binding sites from protein surfaces**. *Nucleic Acids Res* 2012, **40**(Web Server issue):W423-427.

60. Wang L, Liu ZP, Zhang XS, Chen L: **Prediction of hot spots in protein interfaces using a random forest model with hybrid features**. *Protein Eng Des Sel* 2012, **25**(3):119-126.

61. Berggard T, Linse S, James P: **Methods for the detection and analysis of protein-protein interactions**. *Proteomics* 2007, **7**(16):2833-2842.

62. Phizicky EM, Fields S: **Protein-protein interactions: methods for detection and analysis**. *Microbiol Rev* 1995, **59**(1):94-123.

63. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.

64. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.

65. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120.

66. Aloy P, Russell RB: **Ten thousand interactions for the molecular biologist**. *Nat Biotechnol* 2004, **22**(10):1317-1321.

67. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW: **An in vivo map of the yeast protein interactome**. *Science* 2008, **320**(5882):1465-1470.

68. Fields S, Song O: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**(6230):245-246.

69. Remy I, Ghaddar G, Michnick SW: **Using the beta-lactamase protein-fragment complementation assay to probe dynamic protein-protein interactions**. *Nat Protoc* 2007, **2**(9):2302-2306.

70. Cassonnet P, Rolloy C, Neveu G, Vidalain PO, Chantier T, Pellet J, Jones L, Muller M, Demeret C, Gaud G *et al*: **Benchmarking a luciferase complementation assay for detecting protein complexes**. *Nat Methods* 2011, **8**(12):990-992.

71. Dunkler A, Muller J, Johnsson N: **Detecting protein-protein interactions with the Split-Ubiquitin sensor**. *Methods Mol Biol* 2012, **786**:115-130.

72. Johnsson N, Varshavsky A: **Split ubiquitin as a sensor of protein interactions in vivo**. *Proc Natl Acad Sci U S A* 1994, **91**(22):10340-10344.

73. Hazbun TR, Malmstrom L, Anderson S, Graczyk BJ, Fox B, Riffle M, Sundin BA, Aranda JD, McDonald WH, Chiu CH *et al*: **Assigning function to yeast proteins by integration of technologies**. *Molecular cell* 2003, **12**(6):1353-1365.

74. Blakely BT, Rossi FM, Tillotson B, Palmer M, Estelles A, Blau HM: **Epidermal growth factor receptor dimerization monitored in live cells**. *Nat Biotechnol* 2000, **18**(2):218-222.

75. Tafelmeyer P, Johnsson N, Johnsson K: **Transforming a (beta/alpha)8--barrel enzyme into a split-protein sensor through directed evolution**. *Chem Biol* 2004, **11**(5):681-689.

76. Bauch A, Superti-Furga G: **Charting protein complexes, signaling pathways, and networks in the immune system**. *Immunol Rev* 2006, **210**:187-207.

77. Burckstummer T, Bennett KL, Preradovic A, Schutze G, Hantschel O, Superti-Furga G, Bauch A: **An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells**. *Nat Methods* 2006, **3**(12):1013-1019.

78. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B: **A generic protein purification method for protein complex characterization and proteome exploration**. *Nat Biotechnol* 1999, **17**(10):1030-1032.

79. Einarson M: **Detection of protein-protein interactions using the gst fusion protein pulldown technique**. In: *Molecular Cloning: A Laboratory Manual.* 3 edn: Cold Spring Harbor Laboratory Press; 2001.

80. Einarson MB OJ: **Identification of protein-protein interactions with glutathione s-transferase fusion proteins**. In: *Protein-Protein Interactions: A Molecular Cloning Manual.* Cold Spring Harbor Laboratory Press; 2002.

81. Vikis H, Guan, KL: **Glutathione-s-transferase-fusion based assays for studying protein-protein interactions**. In: *Protein-Protein Interactions: Methods and Applications (Methods in Molecular Biology).* Humana Press; 2004.

82. Long F, Cho W, Ishii Y: **Expression and purification of 15N- and 13C-isotope labeled 40-residue human Alzheimer's beta-amyloid peptide for NMR-based structural analysis**. *Protein Expr Purif* 2011, **79**(1):16-24.

83. Benard V, Bokoch GM: **Assay of Cdc42, Rac, and Rho GTPase activation by affinity methods**. *Methods Enzymol* 2002, **345**:349-359.

84. Ren L, Chang E, Makky K, Haas AL, Kaboord B, Walid Qoronfleh M: **Glutathione S-transferase pull-down assays using dehydrated immobilized glutathione resin**. *Anal Biochem* 2003, **322**(2):164-169.

85. al HPe: **Molecular Biomethods Handbook**, 2 edn: Humana Press; 2008.

86. Miernyk JA, Thelen JJ: **Biochemical approaches for discovering protein-protein interactions**. *Plant J* 2008, **53**(4):597-609.

87. Golemis E: **Protein-protein interactions : A molecular cloning manual**. New York: Cold Spring Harbor Laboratory Press; 2002.

88. Persani L, Calebiro D, Bonomi M: **Technology Insight: modern methods to monitor protein-protein interactions reveal functional TSH receptor oligomerization**. *Nat Clin Pract Endocrinol Metab* 2007, **3**(2):180-190.

89. Smith GP: **Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface**. *Science* 1985, **228**(4705):1315-1317.

90. Sundell GN, Ivarsson Y: **Interaction analysis through proteomic phage display**. *Biomed Res Int* 2014, **2014**:176172.

91. Zhang Y, Appleton BA, Wiesmann C, Lau T, Costa M, Hannoush RN, Sidhu SS: **Inhibition of Wnt signaling by Dishevelled PDZ peptides**. *Nat Chem Biol* 2009, **5**(4):217-219.

92. Smith GP, Petrenko VA: **Phage Display**. *Chem Rev* 1997, **97**(2):391-410.

93. Kehoe JW, Kay BK: **Filamentous phage display in the new millennium**. *Chem Rev* 2005, **105**(11):4056-4072.

94. Malys N, Chang DY, Baumann RG, Xie D, Black LW: **A bipartite bacteriophage T4 SOC and HOC randomized peptide display library: detection and analysis of phage T4 terminase (gp17) and late sigma factor (gp55) interaction**. *J Mol Biol* 2002, **319**(2):289-304.

95. Huang R, Pershad, K., Kokoszka, M. and Kay, B. K. (ed.): **Phage-displayed combinatorial peptides, in amino acids, peptides and proteins in organic chemistry**. Weinheim, Germany: WILEY-VCH Verlag GmbH & Co.; 2011.

96. Kay BK, Adey NB, He YS, Manfredi JP, Mataragnon AH, Fowlkes DM: **An M13 phage library displaying random 38-amino-acid peptides as a source of novel sequences with affinity to selected targets**. *Gene* 1993, **128**(1):59-65.

97. Peters EA, Schatz PJ, Johnson SS, Dower WJ: **Membrane insertion defects caused by positive charges in the early mature region of protein pIII of filamentous phage fd can be corrected by prlA suppressors**. *J Bacteriol* 1994, **176**(14):4296-4305.

98. Krumpe LR, Atkinson AJ, Smythers GW, Kandel A, Schumacher KM, McMahon JB, Makowski L, Mori T: **T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries**. *Proteomics* 2006, **6**(15):4210-4222.

99. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**(1):45-48.

100. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TREMBL**. *Nucleic Acids Res* 1996, **24**(1):21-25.

101. Bairoch A: **Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!** *Bioinformatics* 2000, **16**(1):48-64.

102. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R: **High-quality protein knowledge resource: SWISS-PROT and TrEMBL**. *Brief Bioinform* 2002, **3**(3):275-284.

103. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE *et al*: **The Protein Information Resource**. *Nucleic Acids Res* 2003, **31**(1):345-347.

104. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

105. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R: **UniProt archive**. *Bioinformatics* 2004, **20**(17):3236-3237.

106. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics* 2007, **23**(10):1282-1288.

107. **The universal protein resource (UniProt)**. *Nucleic Acids Res* 2008, **36**(Database issue):D190-195.

108. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A *et al*: **neXtProt: a knowledge platform for human proteins**. *Nucleic Acids Res* 2012, **40**(Database issue):D76-83.

109. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

110. Berman HM: **The Protein Data Bank: a historical perspective**. *Acta Crystallogr A* 2008, **64**(Pt 1):88-95.

111. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 2008, **36**(Database issue):D419-425.

112. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA: **Extending CATH: increasing coverage of the protein structure universe and linking structure with function**. *Nucleic Acids Res* 2011, **39**(Database issue):D420-426.

113. Bauer RA, Gunther S, Jansen D, Heeger C, Thaben PF, Preissner R: **SuperSite: dictionary of metabolite and drug binding sites in proteins**. *Nucleic Acids Res* 2009, **37**(Database issue):D195-200.

114. Shin JM, Cho DH: **PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures**. *Nucleic Acids Res* 2005, **33**(Database issue):D238-241.

115. Laskowski RA: **PDBsum new things**. *Nucleic Acids Res* 2009, **37**(Database issue):D355-359.

116. Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F *et al*: **Gene Ontology annotations and resources**. *Nucleic Acids Res* 2013, **41**(Database issue):D530-535.

117. Singh H, Chauhan JS, Gromiha MM, Raghava GP: **ccPDB: compilation and creation of data sets from Protein Data Bank**. *Nucleic Acids Res* 2012, **40**(Database issue):D486-489.

118. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G: **A series of PDB related databases for everyday needs**. *Nucleic Acids Res* 2011, **39**(Database issue):D411-419.

119. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R: **A survey of available tools and web servers for analysis of protein-protein interactions and interfaces**. *Brief Bioinform* 2009, **10**(3):217-232.

120. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res* 2002, **30**(1):303-305.

121. Duan XJ, Xenarios I, Eisenberg D: **Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database**. *Mol Cell Proteomics* 2002, **1**(2):104-116.

122. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D: **Prolinks: a database of protein functional linkages derived from coevolution**. *Genome Biol* 2004, **5**(5):R35.

123. Bader GD, Hogue CW: **BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways**. *Bioinformatics* 2000, **16**(5):465-477.

124. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update**. *Nucleic Acids Res* 2005, **33**(Database issue):D418-424.

125. Isserlin R, El-Badrawi RA, Bader GD: **The Biomolecular Interaction Network Database in PSI-MI 2.5**. *Database (Oxford)* 2011, **2011**:baq037.

126. Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G: **DOMINO: a database of domain-peptide interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D557-560.

127. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database**. *Nucleic Acids Res* 2007, **35**(Database issue):D572-574.

128. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U *et al*: **The IntAct molecular interaction database in 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D841-846.

129. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V *et al*: **The BioGRID Interaction Database: 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D637-640.

130. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

131. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33**(Database issue):D514-517.

132. Szklarczyk D, Franceschini A, Wyder S, Gorslund K, Heller D, Huerte-Cepas J, Simonovic M, Roth A, Santos A, Kalliopi P *et al*: **STRING v10: protein–protein interaction networks, integrated over the tree of life.** *Nucleic Acids Res* 2015, **43**(Database issue):D447-452.

133. Brown KR, Jurisica I: **Online predicted human interaction database**. *Bioinformatics* 2005, **21**(9):2076-2082.

134. Han K, Park B, Kim H, Hong J, Park J: **HPID: the Human Protein Interaction Database**. *Bioinformatics* 2004, **20**(15):2466-2470.

135. McDowall MD, Scott MS, Barton GJ: **PIPs: human protein-protein interaction prediction database**. *Nucleic Acids Res* 2009, **37**(Database issue):D651-656.

136. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30**(1):31-34.

137. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE *et al*: **CYGD: the Comprehensive Yeast Genome Database**. *Nucleic Acids Res* 2005, **33**(Database issue):D364-368.

138. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C *et al*: **The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces**. *Bioinformatics* 2003, **19**(11):1453-1454.

139. Dou Y, Baisnee PF, Pollastri G, Pecout Y, Nowick J, Baldi P: **ICBS: a database of interactions between protein chains mediated by beta-sheet formation**. *Bioinformatics* 2004, **20**(16):2767-2777.

140. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces**. *J Mol Biol* 2003, **325**(2):377-387.

141. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: **Transient protein-protein interactions: structural, functional, and network properties**. *Structure* 2010, **18**(10):1233-1243.

142. Krishna SS, Aravind L: **The bridge-region of the Ku superfamily is an atypical zinc ribbon domain**. *J Struct Biol* 2010, **172**(3):294-299.

143. Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions**. *J Mol Biol* 2003, **325**(5):991-1018.

144. La D, Kong M, Hoffman W, Choi YI, Kihara D: **Predicting permanent and transient protein-protein interfaces**. *Proteins* 2012.

145. Karthikraja V, Suresh A, Lulu S, Kangueane U, Kangueane P: **Types of interfaces for homodimer folding and binding**. *Bioinformation* 2009, **4**(3):101-111.

146. Suresh A, Karthikraja V, Lulu S, Kangueane U, Kangueane P: **A decision tree model for the prediction of homodimer folding mechanism**. *Bioinformation* 2009, **4**(5):197-205.

147. Zhang Z, Witham S, Alexov E: **On the role of electrostatics in protein-protein interactions**. *Phys Biol* 2011, **8**(3):035001.

148. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

149. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect**. *Protein Sci* 1997, **6**(1):53-64.

150. Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites**. *Proteins* 2002, **47**(3):334-343.

151. Brinda KV, Kannan N, Vishveshwara S: **Analysis of homodimeric protein interfaces by graph-spectral methods**. *Protein Eng* 2002, **15**(4):265-277.

152. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces**. *J Mol Biol* 2004, **336**(4):943-955.

153. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13**(1):190-202.

154. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z: **Protein-protein docking benchmark version 3.0**. *Proteins* 2008, **73**(3):705-709.

155. Reynolds C, Damerell D, Jones S: **ProtorP: a protein-protein interaction analysis server**. *Bioinformatics* 2009, **25**(3):413-414.

156. Guharoy M, Chakrabarti P: **Conserved residue clusters at protein-protein interfaces and their use in binding site identification**. *BMC Bioinformatics* 2010, **11**:286.

157. Hwang H, Vreven T, Janin J, Weng Z: **Protein-protein docking benchmark version 4.0**. *Proteins* 2010, **78**(15):3111-3114.

158. Janin J, Rodier F: **Protein-protein interaction at crystal contacts**. *Proteins* 1995, **23**(4):580-587.

159. Dasgupta S, Iyer GH, Bryant SH, Lawrence CE, Bell JA: **Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers**. *Proteins* 1997, **28**(4):494-514.

160. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis**. *J Mol Biol* 1997, **272**(1):133-143.

161. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts**. *J Mol Biol* 2001, **313**(2):399-416.

162. Zhou HX, Shan Y: **Prediction of protein interaction sites from sequence profile and residue neighbor list**. *Proteins* 2001, **44**(3):336-343.

163. Fariselli P, Pazos F, Valencia A, Casadio R: **Prediction of protein--protein interaction sites in heterocomplexes with neural networks**. *Eur J Biochem* 2002, **269**(5):1356-1361.

164. Chen R, Mintseris J, Janin J, Weng Z: **A protein-protein docking benchmark**. *Proteins* 2003, **52**(1):88-91.

165. Mintseris J, Weng Z: **Atomic contact vectors in protein-protein recognition**. *Proteins* 2003, **53**(3):629-639.

166. Zhanhua C, Gan JG, Lei L, Mathura VS, Sakharkar MK, Kangueane P: **Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in euclidian space**. *Front Biosci* 2005, **10**:844-852.

167. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z: **Protein-Protein Docking Benchmark 2.0: an update**. *Proteins* 2005, **60**(2):214-216.

168. Zhu H, Domingues FS, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types**. *BMC Bioinformatics* 2006, **7**:27.

169. Li MH, Lin L, Wang XL, Liu T: **Protein-protein interaction site prediction based on conditional random fields**. *Bioinformatics* 2007, **23**(5):597-604.

170. Keskin O, Nussinov R, Gursoy A: **PRISM: protein-protein interaction prediction by structural matching**. *Methods Mol Biol* 2008, **484**:505-521.

171. Choi YS, Yang JS, Choi Y, Ryu SH, Kim S: **Evolutionary conservation in multiple faces of protein interaction**. *Proteins* 2009, **77**(1):14-25.

172. Gromiha MM, Yokota K, Fukui K: **Energy based approach for understanding the recognition mechanism in protein-protein complexes**. *Mol Biosyst* 2009, **5**(12):1779-1786.

173. Liu R, Jiang W, Zhou Y: **Identifying protein-protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area**. *Amino Acids* 2010, **38**(1):263-270.

174. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J: **A structure-based benchmark for protein-protein binding affinity**. *Protein Sci* 2011, **20**(3):482-491.

175. Swapna LS, Srikeerthana K, Srinivasan N: **Extent of structural asymmetry in homodimeric proteins: prevalence and relevance**. *PLoS One* 2012, **7**(5):e36688.

176. Swapna LS, Mahajan S, de Brevern AG, Srinivasan N: **Comparison of tertiary structures of proteins in protein-protein complexes with unbound forms suggests prevalence of allostery in signalling proteins**. *BMC Struct Biol* 2012, **12**:6.

177. Yugandhar K, Gromiha MM: **Feature selection and classification of protein-protein complexes based on their binding affinities using machine learning approaches**. *Proteins* 2014, **82**(9):2088-2096.

178. Du X, Jing A, Hu X: **A novel feature extraction scheme for prediction of protein-protein interaction sites**. *Mol Biosyst* 2015, **11**(2):475-485.

179. Saha RP, Bahadur RP, Pal A, Mandal S, Chakrabarti P: **ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces**. *BMC Struct Biol* 2006, **6**:11.

180. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility**. *J Mol Biol* 1971, **55**(3):379-400.

181. Bahadur RP, Zacharias M: **The interface of protein-protein complexes: analysis of contacts and prediction of interactions**. *Cell Mol Life Sci* 2008, **65**(7-8):1059-1072.

182. Venkatarajan S Mathura PK: **Protein-protein interaction and macromolecular visualization**. In: *Bioinformatics: A Concept-Based Introduction.* New York: Springer; 2009.

183. Lawrence MC, Colman PM: **Shape complementarity at protein/protein interfaces**. *J Mol Biol* 1993, **234**(4):946-950.

184. Norel R, Lin SL, Wolfson HJ, Nussinov R: **Shape complementarity at protein-protein interfaces**. *Biopolymers* 1994, **34**(7):933-940.

185. Li Y, Zhang X, Cao D: **The role of shape complementarity in the protein-protein interactions**. *Sci Rep* 2013, **3**:3271.

186. Walls PH, Sternberg MJ: **New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking**. *J Mol Biol* 1992, **228**(1):277-297.

187. Tulip WR, Varghese JN, Laver WG, Webster RG, Colman PM: **Refined crystal structure of the influenza virus N9 neuraminidase-NC41 Fab complex**. *J Mol Biol* 1992, **227**(1):122-148.

188. Gabb HA, Jackson RM, Sternberg MJ: **Modelling protein docking using shape complementarity, electrostatics and biochemical information**. *J Mol Biol* 1997, **272**(1):106-120.

189. Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions**. *J Mol Graph* 1995, **13**(5):323-330, 307-328.

190. Gabdoulline RR, Wade RC: **On the protein-protein diffusional encounter complex**. *J Mol Recognit* 1999, **12**(4):226-234.

191. Sheinerman FB, Norel R, Honig B: **Electrostatic aspects of protein-protein interactions**. *Curr Opin Struct Biol* 2000, **10**(2):153-159.

192. McCoy AJ, Chandana Epa V, Colman PM: **Electrostatic complementarity at protein/protein interfaces**. *J Mol Biol* 1997, **268**(2):570-584.

193. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces**. *J Mol Biol* 1998, **280**(1):1-9.

194. Halperin I, Wolfson H, Nussinov R: **Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking**. *Structure* 2004, **12**(6):1027-1038.

195. Ma B, Elkayam T, Wolfson H, Nussinov R: **Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces**. *Proc Natl Acad Sci U S A* 2003, **100**(10):5772-5777.

196. Cukuroglu E, Engin HB, Gursoy A, Keskin O: **Hot spots in protein-protein interfaces: towards drug discovery**. *Prog Biophys Mol Biol* 2014, **116**(2-3):165-173.

197. Keskin O, Ma B, Nussinov R: **Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues**. *J Mol Biol* 2005, **345**(5):1281-1294.

198. Hu Z, Ma B, Wolfson H, Nussinov R: **Conservation of polar residues as hot spots at protein interfaces**. *Proteins* 2000, **39**(4):331-342.

199. Tuncbag N, Gursoy A, Keskin O: **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy**. *Bioinformatics* 2009, **25**(12):1513-1520.

200. Guharoy M, Chakrabarti P: **Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions**. *Bioinformatics* 2007, **23**(15):1909-1918.

201. Gao Y, Wang R, Lai L: **Structure-based method for analyzing protein-protein interfaces**. *J Mol Model* 2004, **10**(1):44-54.

202. Das A, Chakrabarti J, Ghosh M: **Thermodynamics of interfacial changes in a protein-protein complex**. *Mol Biosyst* 2014, **10**(3):437-445.

203. Wang Q, Zhang P, Hoffman L, Tripathi S, Homouz D, Liu Y, Waxham MN, Cheung MS: **Protein recognition and selection through conformational and mutually induced fit**. *Proc Natl Acad Sci U S A* 2013, **110**(51):20545-20550.

204. Marsh JA, Teichmann SA: **Relative solvent accessible surface area predicts protein conformational changes upon binding**. *Structure* 2011, **19**(6):859-867.

205. Ofran Y, Rost B: **ISIS: interaction sites identified from sequence**. *Bioinformatics* 2007, **23**(2):e13-16.

206. Janin J, Miller S, Chothia C: **Surface, subunit interfaces and interior of oligomeric proteins**. *J Mol Biol* 1988, **204**(1):155-164.

207. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N: **Residue frequencies and pairing preferences at protein-protein interfaces**. *Proteins* 2001, **43**(2):89-102.

208. Neuvirth H, Raz R, Schreiber G: **ProMate: a structure based prediction program to identify the location of protein-protein binding sites**. *J Mol Biol* 2004, **338**(1):181-199.

209. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15447-15452.

210. Gromiha MM, Saranya N, Selvaraj S, Jayaram B, Fukui K: **Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes**. *Proteome Sci* 2011, **9 Suppl 1**:S13.

211. Mosca R, Ceol A, Aloy P: **Interactome3D: adding structural details to protein networks**. *Nat Methods* 2012, **10**(1):47-53.

212. Davis FP, Sali A: **PIBASE: a comprehensive database of structurally defined protein interfaces**. *Bioinformatics* 2005, **21**(9):1901-1907.

213. Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces**. *Nucleic Acids Res* 2006, **34**(Database issue):D310-314.

214. Stein A, Russell RB, Aloy P: **3did: interacting protein domains of known three-dimensional structure**. *Nucleic Acids Res* 2005, **33**(Database issue):D413-417.

215. Mosca R, Ceol A, Stein A, Olivella R, Aloy P: **3did: a catalog of domain-based interactions of known three-dimensional structure**. *Nucleic Acids Res* 2014, **42**(Database issue):D374-379.

216. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends Biochem Sci* 1998, **23**(9):358-361.

217. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

218. Knudsen M, Wiuf C: **The CATH database**. *Hum Genomics* 2010, **4**(3):207-212.

219. Laskowski RA: **PDBsum: summaries and analyses of PDB structures**. *Nucleic Acids Res* 2001, **29**(1):221-222.

220. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

221. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Res* 2001, **29**(1):37-40.

222. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R *et al*: **New developments in the InterPro database**. *Nucleic Acids Res* 2007, **35**(Database issue):D224-228.

223. Teyra J, Doms A, Schroeder M, Pisabarro MT: **SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces**. *BMC Bioinformatics* 2006, **7**:104.

224. Kundrotas PJ, Alexov E: **PROTCOM: searchable database of protein complexes enhanced with domain-domain structures**. *Nucleic Acids Res* 2007, **35**(Database issue):D575-579.

225. Kumar MD, Gromiha MM: **PINT: Protein-protein Interactions Thermodynamic Database**. *Nucleic Acids Res* 2006, **34**(Database issue):D195-198.

226. Aloy P, Russell RB: **InterPreTS: protein interaction prediction through tertiary structure**. *Bioinformatics* 2003, **19**(1):161-162.

227. Davis FP, Braberg H, Shen MY, Pieper U, Sali A, Madhusudhan MS: **Protein complex compositions predicted by structural similarity**. *Nucleic Acids Res* 2006, **34**(10):2943-2952.

228. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D *et al*: **MODBASE: a database of annotated comparative protein structure models and associated resources**. *Nucleic Acids Res* 2006, **34**(Database issue):D291-295.

229. Murakami Y, Jones S: **SHARP2: protein-protein interaction predictions using patch analysis**. *Bioinformatics* 2006, **22**(14):1794-1795.

230. Qin S, Zhou HX: **meta-PPISP: a meta web server for protein-protein interaction site prediction**. *Bioinformatics* 2007, **23**(24):3386-3387.

231. Tjong H, Qin S, Zhou HX: **PI2PE: protein interface/interior prediction engine**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W357-362.

232. Tina KG, Bhadra R, Srinivasan N: **PIC: Protein Interactions Calculator**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W473-476.

233. Wallace AC, Laskowski RA, Thornton JM: **LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions**. *Protein Eng* 1995, **8**(2):127-134.

234. Bradford JR, Westhead DR: **Improved prediction of protein-protein binding sites using a support vector machines approach**. *Bioinformatics* 2005, **21**(8):1487-1494.

235. Reulecke I, Lange G, Albrecht J, Klein R, Rarey M: **Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function**. *ChemMedChem* 2008, **3**(6):885-897.

236. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ: **Spatial chemical conservation of hot spot interactions in protein-protein complexes**. *BMC Biol* 2007, **5**:43.

237. de Vries SJ, van Dijk AD, Bonvin AM: **WHISCY: what information does surface conservation yield? Application to data-driven docking**. *Proteins* 2006, **63**(3):479-489.

238. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ: **SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D580-589.

239. Liang S, Zhang C, Liu S, Zhou Y: **Protein binding site prediction using an empirical scoring function**. *Nucleic Acids Res* 2006, **34**(13):3698-3707.

240. Chang DT, Weng YZ, Lin JH, Hwang MJ, Oyang YJ: **Protemot: prediction of protein binding sites with automatically extracted geometrical templates**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W303-309.

241. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W299-302.

242. Negi SS, Schein CH, Oezguen N, Power TD, Braun W: **InterProSurf: a web server for predicting interacting sites on protein surfaces**. *Bioinformatics* 2007, **23**(24):3397-3399.

243. Guharoy M, Pal A, Dasgupta M, Chakrabarti P: **PRICE (PRotein Interface Conservation and Energetics): a server for the analysis of protein-protein interfaces**. *J Struct Funct Genomics* 2011, **12**(1):33-41.

244. Yugandhar K, Gromiha MM: **Protein-protein binding affinity prediction from amino acid sequence**. *Bioinformatics* 2014, **30**(24):3583-3589.

245. Orii N, Ganapathiraju MK: **Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function**. *PLoS One* 2012, **7**(11):e49029.

246. Tuncbag N, Keskin O, Gursoy A: **HotPoint: hot spot prediction server for protein interfaces**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W402-406.

247. Pavelka A, Chovancova E, Damborsky J: **HotSpot Wizard: a web server for identification of hot spots in protein engineering**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W376-383.

248. Lise S, Buchan D, Pontil M, Jones DT: **Predictions of hot spot residues at protein-protein interfaces using support vector machines**. *PLoS One* 2011, **6**(2):e16774.

249. Darnell SJ, LeGault L, Mitchell JC: **KFC Server: interactive forecasting of protein interaction hot spots**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W265-269.

250. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S: **PredHS: a web server for predicting protein-protein interaction hot spots by using structural neighborhood properties**. *Nucleic Acids Res* 2014, **42**(Web Server issue):W290-295.

251. Thorn KS, Bogan AA: **ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions**. *Bioinformatics* 2001, **17**(3):284-285.

252. Guney E, Tuncbag N, Keskin O, Gursoy A: **HotSprint: database of computational hot spots in protein interfaces**. *Nucleic Acids Res* 2008, **36**(Database issue):D662-666.

253. Xia JF, Zhao XM, Song J, Huang DS: **APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility**. *BMC Bioinformatics* 2010, **11**:174.

254. Li Z, Wong L, Li J: **DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts**. *BMC Syst Biol* 2011, **5 Suppl 1**:S5.

255. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ: **CAPRI: a Critical Assessment of PRedicted Interactions**. *Proteins* 2003, **52**(1):2-9.

256. Janin J: **Assessing predictions of protein-protein interaction: the CAPRI experiment**. *Protein Sci* 2005, **14**(2):278-283.

257. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G *et al*: **Community-wide assessment of protein-interface modeling suggests improvements to design methodology**. *J Mol Biol* 2011, **414**(2):289-302.

258. Kim C, Ye F, Ginsberg MH: **Regulation of integrin activation**. *Annu Rev Cell Dev Biol* 2011, **27**:321-345.

259. Humphries MJ: **Integrin structure**. *Biochem Soc Trans* 2000, **28**(4):311-339.

260. Hynes RO: **The emergence of integrins: a personal and historical perspective**. *Matrix Biol* 2004, **23**(6):333-340.

261. Xiong JP, Stehle T, Diefenbach B, Zhang R, Dunker R, Scott DL, Joachimiak A, Goodman SL, Arnaout MA: **Crystal structure of the extracellular segment of integrin alpha Vbeta3**. *Science* 2001, **294**(5541):339-345.

262. Xiong JP, Stehle T, Zhang R, Joachimiak A, Frech M, Goodman SL, Arnaout MA: **Crystal structure of the extracellular segment of integrin alpha Vbeta3 in complex with an Arg-Gly-Asp ligand**. *Science* 2002, **296**(5565):151-155.

263. Giancotti FG, Ruoslahti E: **Integrin signaling**. *Science* 1999, **285**(5430):1028-1032.

264. Soung YH, Clifford JL, Chung J: **Crosstalk between integrin and receptor tyrosine kinase signaling in breast carcinoma progression**. *BMB Rep* 2010, **43**(5):311-318.

265. Arnaout MA, Goodman SL, Xiong JP: **Structure and mechanics of integrin-based cell adhesion**. *Curr Opin Cell Biol* 2007, **19**(5):495-507.

266. Hynes RO: **Integrins: bidirectional, allosteric signaling machines**. *Cell* 2002, **110**(6):673-687.

267. Agrez M, Chen A, Cone RI, Pytela R, Sheppard D: **The alpha v beta 6 integrin promotes proliferation of colon carcinoma cells through a unique region of the beta 6 cytoplasmic domain**. *J Cell Biol* 1994, **127**(2):547-556.

268. Ramos DM, Dang D, Sadler S: **The role of the integrin alpha v beta6 in regulating the epithelial to mesenchymal transition in oral cancer**. *Anticancer Res* 2009, **29**(1):125-130.

269. Hazelbag S, Kenter GG, Gorter A, Dreef EJ, Koopman LA, Violette SM, Weinreb PH, Fleuren GJ: **Overexpression of the alpha v beta 6 integrin in cervical squamous cell carcinoma is a prognostic factor for decreased survival**. *J Pathol* 2007, **212**(3):316-324.

270. Lu X, Lu D, Scully M, Kakkar V: **The role of integrins in cancer and the development of anti-integrin therapeutic agents for cancer therapy**. *Perspect Medicin Chem* 2008, **2**:57-73.

271. Van Aarsen LA, Leone DR, Ho S, Dolinski BM, McCoon PE, LePage DJ, Kelly R, Heaney G, Rayhorn P, Reid C *et al*: **Antibody-mediated blockade of integrin alpha v beta 6 inhibits tumor progression in vivo by a transforming growth factor-beta-regulated mechanism**. *Cancer Res* 2008, **68**(2):561-570.

272. Binder M, Trepel M: **Drugs targeting integrins for cancer therapy**. *Expert Opin Drug Dis* 2009, **4**(3):229-241.

273. Margadant C, Sonnenberg A: **Integrin-TGF-beta crosstalk in fibrosis, cancer and wound healing**. *EMBO Rep* 2010, **11**(2):97-105.

274. Campbell ID, Humphries MJ: **Integrin structure, activation, and interactions**. *Cold Spring Harb Perspect Biol* 2011, **3**(3).

275. Barczyk M, Carracedo S, Gullberg D: **Integrins**. *Cell Tissue Res* 2010, **339**(1):269-280.

276. Takada Y, Ye X, Simon S: **The integrins**. *Genome Biol* 2007, **8**(5):215.

277. Green LJ, Mould AP, Humphries MJ: **The integrin beta subunit**. *Int J Biochem Cell Biol* 1998, **30**(2):179-184.

278. Mould AP: **Getting integrins into shape: recent insights into how integrin activity is regulated by conformational changes**. *J Cell Sci* 1996, **109 ( Pt 11)**:2613-2618.

279. Xiong JP, Mahalingham B, Alonso JL, Borrelli LA, Rui X, Anand S, Hyman BT, Rysiok T, Muller-Pompalla D, Goodman SL *et al*: **Crystal structure of the complete integrin alphaVbeta3 ectodomain plus an alpha/beta transmembrane fragment**. *J Cell Biol* 2009, **186**(4):589-600.

280. Bandyopadhyay A, Raghavan S: **Defining the role of integrin alphavbeta6 in cancer**. *Curr Drug Targets* 2009, **10**(7):645-652.

281. Busk M, Pytela R, Sheppard D: **Characterization of the integrin alpha v beta 6 as a fibronectin-binding protein**. *J Biol Chem* 1992, **267**(9):5790-5796.

282. Bates RC, Bellovin DI, Brown C, Maynard E, Wu B, Kawakatsu H, Sheppard D, Oettgen P, Mercurio AM: **Transcriptional activation of integrin beta6 during the epithelial-mesenchymal transition defines a novel prognostic indicator of aggressive colon carcinoma**. *J Clin Invest* 2005, **115**(2):339-347.

283. Breuss JM, Gallo J, DeLisser HM, Klimanskaya IV, Folkesson HG, Pittet JF, Nishimura SL, Aldape K, Landers DV, Carpenter W *et al*: **Expression of the beta 6 integrin subunit in development, neoplasia and tissue repair suggests a role in epithelial remodeling**. *J Cell Sci* 1995, **108 ( Pt 6)**:2241-2251.

284. AlDahlawi S, Eslami A, Hakkinen L, Larjava HS: **The alphavbeta6 integrin plays a role in compromised epidermal wound healing**. *Wound Repair Regen* 2006, **14**(3):289-297.

285. Morris DG, Huang X, Kaminski N, Wang Y, Shapiro SD, Dolganov G, Glick A, Sheppard D: **Loss of integrin alpha(v)beta6-mediated TGF-beta activation causes Mmp12-dependent emphysema**. *Nature* 2003, **422**(6928):169-173.

286. Munger JS, Huang X, Kawakatsu H, Griffiths MJ, Dalton SL, Wu J, Pittet JF, Kaminski N, Garat C, Matthay MA *et al*: **The integrin alpha v beta 6 binds and activates latent TGF beta 1: a mechanism for regulating pulmonary inflammation and fibrosis**. *Cell* 1999, **96**(3):319-328.

287. Ramsay AG, Keppler MD, Jazayeri M, Thomas GJ, Parsons M, Violette S, Weinreb P, Hart IR, Marshall JF: **HS1-associated protein X-1 regulates carcinoma cell migration and invasion via clathrin-mediated endocytosis of integrin alphavbeta6**. *Cancer research* 2007, **67**(11):5275-5284.

288. Lanzetti L, Di Fiore PP: **Endocytosis and cancer: an 'insider' network with dangerous liaisons**. *Traffic* 2008, **9**(12):2011-2021.

289. Morgan MR, Thomas GJ, Russell A, Hart IR, Marshall JF: **The integrin cytoplasmic-tail motif EKQKVDLSTDC is sufficient to promote tumor cell invasion mediated by matrix metalloproteinase (MMP)-2 or MMP-9**. *J Biol Chem* 2004, **279**(25):26533-26539.

290. Diamond MS, Springer TA: **The dynamic regulation of integrin adhesiveness**. *Curr Biol* 1994, **4**(6):506-517.

291. Sheppard D, Rozzo C, Starr L, Quaranta V, Erle DJ, Pytela R: **Complete amino acid sequence of a novel integrin beta subunit (beta 6) identified in epithelial cells using the polymerase chain reaction**. *J Biol Chem* 1990, **265**(20):11502-11507.

292. Yang SB, Du Y, Wu BY, Xu SP, Wen JB, Zhu M, Cai CH, Yang PC: **Integrin alphavbeta6 promotes tumor tolerance in colorectal cancer**. *Cancer Immunol Immunother* 2012, **61**(3):335-342.

293. Hamidi S, Salo T, Kainulainen T, Epstein J, Lerner K, Larjava H: **Expression of alpha(v)beta6 integrin in oral leukoplakia**. *Br J Cancer* 2000, **82**(8):1433-1440.

294. Ahmed N, Pansino F, Clyde R, Murthi P, Quinn MA, Rice GE, Agrez MV, Mok S, Baker MS: **Overexpression of alpha(v)beta6 integrin in serous epithelial ovarian cancer regulates extracellular matrix degradation via the plasminogen activation cascade**. *Carcinogenesis* 2002, **23**(2):237-244.

295. Ahmed N, Riley C, Rice GE, Quinn MA, Baker MS: **Alpha(v)beta(6) integrin-A marker for the malignant potential of epithelial ovarian cancer**. *J Histochem Cytochem* 2002, **50**(10):1371-1380.

296. Arihiro K, Kaneko M, Fujii S, Inai K, Yokosaki Y: **Significance of alpha 9 beta 1 and alpha v beta 6 integrin expression in breast carcinoma**. *Breast Cancer* 2000, **7**(1):19-26.

297. Breuss JM, Gillett N, Lu L, Sheppard D, Pytela R: **Restricted distribution of integrin beta 6 mRNA in primate epithelial tissues**. *J Histochem Cytochem* 1993, **41**(10):1521-1527.

298. Hazelbag S, Kenter GG, Gorter A, Dreef EJ, Koopman LA, Violette SM, Weinreb PH, Fleuren GJ: **Overexpression of the alpha v beta 6 integrin in cervical squamous cell carcinoma is a prognostic factor for decreased survival**. *J Pathol* 2007, **212**(3):316-324.

299. Hecht JL, Dolinski BM, Gardner HA, Violette SM, Weinreb PH: **Overexpression of the alphavbeta6 integrin in endometrial cancer**. *Appl Immunohistochem Mol Morphol* 2008, **16**(6):543-547.

300. Kawashima A, Tsugawa S, Boku A, Kobayashi M, Minamoto T, Nakanishi I, Oda Y: **Expression of alphav integrin family in gastric carcinomas: increased alphavbeta6 is associated with lymph node metastasis**. *Pathol Res Pract* 2003, **199**(2):57-64.

301. Marsh D, Dickinson S, Neill GW, Marshall JF, Hart IR, Thomas GJ: **alpha vbeta 6 Integrin promotes the invasion of morphoeic basal cell carcinoma through stromal modulation**. *Cancer Res* 2008, **68**(9):3295-3303.

302. Xue H, Atakilit A, Zhu W, Li X, Ramos DM, Pytela R: **Role of the alpha(v)beta6 integrin in human oral squamous cell carcinoma growth in vivo and in vitro**. *Biochem Biophys Res Commun* 2001, **288**(3):610-618.

303. Ramos DM, But M, Regezi J, Schmidt BL, Atakilit A, Dang D, Ellis D, Jordan R, Li X: **Expression of integrin beta 6 enhances invasive behavior in oral squamous cell carcinoma**. *Matrix Biol* 2002, **21**(3):297-307.

304. Regezi JA, Ramos DM, Pytela R, Dekker NP, Jordan RC: **Tenascin and beta 6 integrin are overexpressed in floor of mouth in situ carcinomas and invasive squamous cell carcinomas**. *Oral Oncol* 2002, **38**(4):332-336.

305. Sipos B, Hahn D, Carceller A, Piulats J, Hedderich J, Kalthoff H, Goodman SL, Kosmahl M, Kloppel G: **Immunohistochemical screening for beta6-integrin subunit expression in adenocarcinomas using a novel monoclonal antibody reveals strong up-regulation in pancreatic ductal adenocarcinomas in vivo and in vitro**. *Histopathology* 2004, **45**(3):226-236.

306. Yang GY, Xu KS, Pan ZQ, Zhang ZY, Mi YT, Wang JS, Chen R, Niu J: **Integrin alpha v beta 6 mediates the potential for colon cancer cells to colonize in and metastasize to the liver**. *Cancer Sci* 2008, **99**(5):879-887.

307. Zhang ZY, Xu KS, Wang JS, Yang GY, Wang W, Wang JY, Niu WB, Liu EY, Mi YT, Niu J: **Integrin alphanvbeta6 acts as a prognostic indicator in gastric carcinoma**. *Clin Oncol (R Coll Radiol)* 2008, **20**(1):61-66.

308. Stoppelli MP, Corti A, Soffientini A, Cassani G, Blasi F, Assoian RK: **Differentiation-enhanced binding of the amino-terminal fragment of human urokinase plasminogen activator to a specific receptor on U937 monocytes**. *Proceedings of the National Academy of Sciences of the United States of America* 1985, **82**(15):4939-4943.

309. Borglum AD, Byskov A, Ragno P, Roldan AL, Tripputi P, Cassani G, Dano K, Blasi F, Bolund L, Kruse TA: **Assignment of the urokinase-type plasminogen activator receptor gene (PLAUR) to chromosome 19q13.1-q13.2**. *Am J Hum Genet* 1992, **50**(3):492-497.

310. Blasi F, Carmeliet P: **uPAR: a versatile signalling orchestrator**. *Nat Rev Mol Cell Biol* 2002, **3**(12):932-943.

311. Pyke C, Graem N, Ralfkiaer E, Ronne E, Hoyer-Hansen G, Brunner N, Dano K: **Receptor for urokinase is present in tumor-associated macrophages in ductal breast carcinoma**. *Cancer research* 1993, **53**(8):1911-1915.

312. Baker MS, Liang XM, Doe WF: **Occupancy of the cancer cell urokinase receptor (uPAR): effects of acid elution and exogenous uPA on cell surface urokinase (uPA)**. *Biochim Biophys Acta* 1992, **1117**(2):143-152.

313. Nykjaer A, Moller B, Todd RF, 3rd, Christensen T, Andreasen PA, Gliemann J, Petersen CM: **Urokinase receptor. An activation antigen in human T lymphocytes**. *J Immunol* 1994, **152**(2):505-516.

314. Ellis V, Behrendt N, Dano K: **Plasminogen activation by receptor-bound urokinase. A kinetic study with both cell-associated and isolated receptor**. *The Journal of biological chemistry* 1991, **266**(19):12752-12758.

315. Lanza F, Castoldi GL, Castagnari B, Todd RF, 3rd, Moretti S, Spisani S, Latorraca A, Focarile E, Roberti MG, Traniello S: **Expression and functional role of urokinase-type plasminogen activator receptor in normal and acute leukaemic cells**. *Br J Haematol* 1998, **103**(1):110-123.

316. Morita S, Sato A, Hayakawa H, Ihara H, Urano T, Takada Y, Takada A: **Cancer cells overexpress mRNA of urokinase-type plasminogen activator, its receptor and inhibitors in human non-small-cell lung cancer tissue: analysis by Northern blotting and in situ hybridization**. *Int J Cancer* 1998, **78**(3):286-292.

317. Mekkawy AH, Morris DL, Pourgholami MH: **Urokinase plasminogen activator system as a potential target for cancer therapy**. *Future Oncol* 2009, **5**(9):1487-1499.

318. Smith HW, Marshall CJ: **Regulation of cell signalling by uPAR**. *Nat Rev Mol Cell Biol* 2010, **11**(1):23-36.

319. Llinas P, Le Du MH, Gardsvoll H, Dano K, Ploug M, Gilquin B, Stura EA, Menez A: **Crystal structure of the human urokinase plasminogen activator receptor bound to an antagonist peptide**. *EMBO J* 2005, **24**(9):1655-1663.

320. Llinas P, Le Du MH, Gardsvoll H, Dano K, Ploug M, Gilquin B, Stura EA, Menez A: **Crystal structure of the human urokinase plasminogen activator receptor bound to an antagonist peptide**. *The EMBO journal* 2005, **24**(9):1655-1663.

321. Eden G, Archinti M, Furlan F, Murphy R, Degryse B: **The urokinase receptor interactome**. *Curr Pharm Des* 2011, **17**(19):1874-1889.

322. Saldanha RG, Molloy MP, Bdeir K, Cines DB, Song X, Uitto PM, Weinreb PH, Violette SM, Baker MS: **Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy**. *J Proteome Res* 2007, **6**(3):1016-1028.

323. Sanchez R, Sali A: **Advances in comparative protein-structure modelling**. *Curr Opin Struct Biol* 1997, **7**(2):206-214.

324. Sowmya G, Khan JM, Anand S, Ahn SB, Baker MS, Ranganathan S: **A site for direct integrin alphavbeta6.uPAR interaction from structural modelling and docking**. *J Struct Biol* 2014, **185**(3):327-335.

325. Chaurasia P, Mezei M, Zhou MM, Ossowski L: **Computer aided identification of small molecules disrupting uPAR/alpha5beta1--integrin interaction: a new paradigm for metastasis prevention**. *PLoS One* 2009, **4**(2):e4617.

326. Laraia L, McKenzie G, Spring DR3, Venkitaraman AR, Huggins DJ: **Overcoming chemical, biological, and computational challenges in the development of inhibitors targeting protein-protein interactions.** *Chem Biol* 2015, **22**:689-703.

# Appendix 1 - Integrin αvβ6·uPAR interactions using functional proteomics

### *Publication 5*

*Reproduced with permission from [Ahn SB, Mohamedali A, Anand S, Cheruku HR, Birch D, Sowmya G, Cantor D, Ranganathan S, Inglis DW, Frank R, Agrez M, Nice EC, Baker MS.* ***Characterisation of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics*** *(2014), Journal of Proteome Research 13(12):5956-64] appendix*

Journal of proteome research

Article

pubs.acs.org/jpr

# Characterization of the Interaction between Heterodimeric $\alpha v\beta 6$ Integrin and Urokinase Plasminogen Activator Receptor (uPAR) Using Functional Proteomics

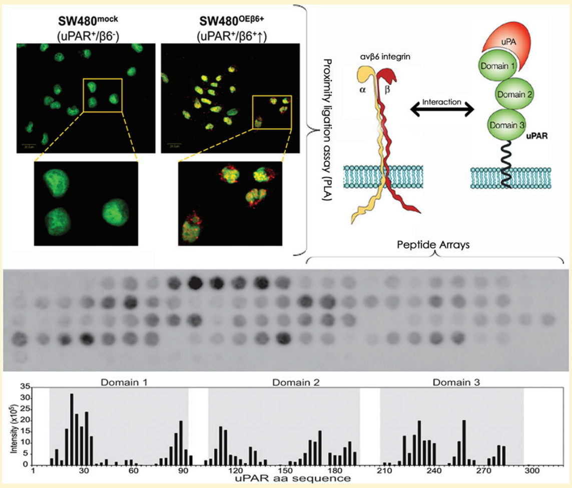
Seong Beom Ahn,[†,∇] Abidali Mohamedali,[‡,∇] Samyuktha Anand,[‡,∇] Harish R. Cheruku,[†] Debra Birch,[‡] Gopichandran Sowmya,[‡] David Cantor,[†] Shoba Ranganathan,[‡] David W. Inglis,[§] Ronald Frank,[||] Michael Agrez,[⊥] Edouard C. Nice,[#] and Mark S. Baker*[,†]

[†]Australian School of Advanced Medicine, Faculty of Human Sciences, [‡]Department of Chemistry and Biomolecular Sciences, and [§]Department of Engineering, Faculty of Science, Macquarie University, Sydney, NSW 2109, Australia

[||]Department of Chemical Biology, Helmholtz Centre for Infection Research, Inhoffen Strasse, 738124 Braunschweig, Germany

[⊥]Division of Surgery, John Hunter Hospital, Newcastle, NSW 2310, Australia

[#]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia



**ABSTRACT:** Urokinase plasminogen activator receptor (uPAR) and the epithelial integrin $\alpha v\beta 6$ are thought to individually play critical roles in cancer metastasis. These observations have been highlighted by the recent discovery (by proteomics) of an interaction between these two molecules, which are also both implicated in the epithelial−mesenchymal transition (EMT) that facilitates escape of cells from tissue barriers and is a common signature of cancer metastases. In this study, orthogonal in cellulo and in vitro functional proteomic approaches were used to better characterize the uPAR·$\alpha v\beta 6$ interaction. Proximity ligation assays (PLA) confirmed the uPAR·$\alpha v\beta 6$ interaction on OVCA429 (ovarian cancer line) and four different colon cancer cell lines including positive controls in cells with de novo $\beta 6$ subunit expression. PLA studies were then validated using peptide arrays, which also identified potential physical sites of uPAR interaction with $\alpha v\beta 6$, as well as verifying interactions with other known uPAR ligands (e.g., uPA, vitronectin) and individual integrin subunits (i.e., $\alpha v$, $\beta 1$, $\beta 3$, and $\beta 6$ alone). Our data suggest that interaction with uPAR requires expression of the complete $\alpha\beta$ heterodimer (e.g., $\alpha v\beta 6$), not individual subunits (i.e., $\alpha v$, $\beta 1$, $\beta 3$, or $\beta 6$). Finally, using in silico structural analyses in concert with these functional proteomics studies, we propose and demonstrate that the most likely unique sites of interaction between $\alpha v\beta 6$ and uPAR are located in uPAR domains II and III.

**KEYWORDS:** *functional proteomics, uPAR, αvβ6 integrin, proximity ligation assay, peptide array, ovarian cancer, colorectal cancer*

## ■ INTRODUCTION

A hallmark of epithelial cancer metastasis is the ability of cancer cells to migrate and infiltrate distant organs. Key stages during metastasis include detachment of the tumor cell from neighboring cells and the basement membrane, intravasation of cell(s) to the blood or lymphatic system, invasion of the migrated cell into a new environment, readhesion, and finally angiogenesis.[1] At the molecular level, the epithelial−mesenchymal transition (EMT) is thought to be a pivotal biological process that facilitates tissue remodeling and metastatic progression. Normal epithelial cells undergo numerous biochemical alterations during EMT, including loss of cell polarity, loss of cell−cell adhesion, suppression of E-cadherin, and an increase in cell migration and invasiveness.[2] EMT is facilitated by degradation of extracellular matrix (ECM) structures, allowing cancer cells to escape and potentially colonize secondary sites in the body.[2] Degradation of ECM is now thought to be one of the most complex and important mechanisms that drives EMT, but how this occurs is not yet fully understood. The matrix metalloproteinase (MMP) family and the serine protease plasminogen activation cascade are two major matrix degrading protease families implicated in epithelial cancer metastasis (e.g., breast, endometrial, hepato-

**Received:** August 15, 2014
**Published:** October 16, 2014

cellular, colorectal, pancreatic, gastric, renal, brain, and lung).[3] Both the MMPs and the plasmin are found as inactive zymogens (pro-MMPs and plasminogen, respectively), which are spatially and temporally (spatiotemporally) activated in a series of steps.[4] Inactive plasminogen can be converted to active plasmin by urokinase plasminogen activator (uPA) on its major receptor the uPA-receptor (uPAR), where it is relatively "shielded" from inhibitors when located on the cell surface. Plasmin degrades many ECM components including fibrin, fibronectin, laminin, and the protein core of proteoglycans,[4] while also activating MMP-1, MMP-3, and MMP-9 among many proteases that consequently degrade additional ECM components.[3] To understand the regulation and consequences of ECM degradation in the tumor microenvironment, it was essential to determine cell surface interacting proteins. Using immunoprecipitation and mass spectrometry, we recently elucidated a cell surface uPAR interactome using an ovarian cancer cell line (OVCA429) with the novel discovery of the interaction of uPAR and integrin $\alpha v\beta 6$,[5] subsequently shown as uPAR·$\alpha v\beta 6$. This was further validated by Western blot analysis. Interestingly, both of these cell surface proteins have been implicated in many aspects of the biology of epithelial cancer and its progression.[5]

From more than 8000 membrane proteins predicted from the human protein-coding genes,[6] uPAR has been suggested to be one of a few multifunctional multi-interacting cell surface receptors that is known to be involved in, among other things, ECM degradation, growth factor activation, and downstream cellular signaling.[7] A glycosylphosphotidylinositol (GPI) linker anchors the three domains (DI, DII, and DIII) of the mature uPAR protein to the extracellular surface of the plasma membrane. These three domains form a thick-fingered glove-like structure that provides a central pocket for the binding of the cognate ligand protease, uPA.[8] Equally this shape reveals a large contralateral external surface potentially facilitating interactions with other proteins.[8] While initial studies focused exclusively on regulation of plasmin activation by uPAR, 42 proteins (9 extracellular proteins and 33 lateral interacting partners) have now been proposed to interact with uPAR.[9] This exhaustive list suggests that uPAR may have evolved multiple different ligand specificities involved in the regulation of many biologies, like proteolysis, cell migration, proliferation, cell signaling, as well as other yet to be explored cell behaviors. Indeed, in the past decade, extensive evidence has suggested that uPAR is implicated in cell adhesion, proliferation, migration, tissue remodeling, and regulation of signaling pathways (e.g., MAP kinase, Ras pathways),[7] which are important features not only of ubiquitous developmental pathways, but more importantly for cancer metastasis. High expression of the uPAR antigen has been observed in many cancers (including breast, ovarian, colon, and lung[10,11]). In colorectal cancer (CRC), a high level of uPAR has been suggested as a prognostic factor for poor survival.[11] Additionally, up-regulation of uPA in metastasis and its subsequent roles in the degradation of the ECM have further suggested uPAR and its interacting partners are central to processes that lead to metastasis, including EMT.[12]

As uPAR possesses no intrinsic intracellular domain, it is commonly thought that downstream cellular signaling pathways influenced by uPAR must be mediated through lateral interactions with transmembrane proteins (e.g., integrins). Indeed, 11 integrins (out of a total of 24) have been suggested to directly interact with uPAR,[9] and many of these studies have implicated these interactions in some role in cancer metastasis.[13] A major function of integrins that relates them directly to cell adhesion in cancer metastasis is in cellular traction, where the $\beta$ subunit embeds itself across the cell membrane and mechanically links integrins to the cytoskeleton and ECM.[13] Integrins also regulate molecular processes related to cell morphology, proliferation, survival, migration, and invasion, mostly by engaging in crucial intracellular signaling.[13]

This study focuses specifically on the $\alpha v\beta 6$ integrin, a transmembrane heterodimer receptor expressed exclusively on the surface of epithelial cells. The $\alpha v\beta 6$ integrin is involved in a bidirectional manner in the signal cascade system, sending signals from the cells to the ECM and vice versa via a series of protein binding partners, which include fibronectin, cytotactin, tenascin, vitronectin (Vn), and TGF$\beta$1.[14] High expression of $\alpha v\beta 6$ has been demonstrated in various cancers including CRC, liver, ovarian, gastric, thyroid, cervical squamous, and endometrial cancer, where its expression is often correlated with poor patient survival.[15,16] Several studies have implicated $\alpha v\beta 6$ in cell proliferation, migration, and invasion,[16,17] with some reports suggesting the involvement of $\alpha v\beta 6$ through activation and up-regulation of various MMP-driven proteolytic pathways.[16] Furthermore, it has been conclusively demonstrated that $\alpha v\beta 6$ activates nascent latent transforming growth factor, TGF-$\beta$1,[18] which can also up-regulate MMP pathways,[19] leading to similar outcomes.

Our central hypothesis here is that, when coexpressed, uPAR and $\alpha v\beta 6$ function cooperatively as a single membrane proteomic machine (as uPAR·$\alpha v\beta 6$). In this study, we confirm the originally observed uPAR·$\alpha v\beta 6$ interaction by functional proteomics using two orthogonal techniques, proximity ligation assays (PLA) and peptide arrays. In detail, PLA is an in cellulo technique that allows direct detection of protein−protein interactions due to the close proximity of the binding partners, and the in vitro peptide array method was used to locate potential specific interacting sites in uPAR·$\alpha v\beta 6$ using an offset 15-mer sequential array of uPAR peptides across the whole protein sequence to find binding sites using HRP-labeled $\alpha v\beta 6$ or other ligands (i.e., uPA, Vn, and integrin subunits). Furthermore, using an in silico structural analysis tool (ICM bioinformatics software), we were able to map putative sites of uPAR and $\alpha v\beta 6$ interaction. This study not only validates the uPAR·$\alpha v\beta 6$ interactions observed by proteomics in CRC and ovarian cancer cells, but also opens significant new avenues for functional targeting of similar interactions that may play key roles in epithelial cancer metastasis and provide unique therapeutic options.

## ■ MATERIALS AND METHODS

### Antibodies and Recombinant Proteins

Monoclonal antibodies (mAb) against human uPAR (clone R4, IgG1) were purchased from DAKO (Glostrup, Denmark). The mAb against the $\beta$6 subunit of the human $\alpha v\beta 6$ integrin (clone 6.4B4, IgG1) was obtained from Biogen Idec (Cambridge, MA).[20] Isotype control, IgG1, was purchased from R&D Systems (Minneapolis, MN). The full length recombinant proteins that were used for the peptide array were uPA and integrin $\alpha v\beta 6$ (R&D Systems); vitronectin (Merck Millipore, MA); and integrin $\alpha$v, $\beta$6, $\beta$1, and $\beta$3 (Abnova, Taipei City, Taiwan).

### Cell Culture

The ovarian and colon cancer cell lines expressing uPAR and varying levels of $\beta 6$ used for the experiments were: ovarian, OVCA429[21] (uPAR$^+$, $\beta 6^+$); colorectal, HT29$^{\mathrm{mock}}$ (uPAR$^+$, $\beta 6^+$), HT29$^{\beta 6\mathrm{AS}}$ (uPAR$^+$, $\beta 6^+\downarrow$), SW480$^{\beta 6\mathrm{OE}}$ (uPAR$^+$, $\beta 6^+\uparrow$), and SW480$^{\mathrm{mock}}$ (uPAR$^+$, $\beta 6^-$).[22,23] The OVCA429 cells were cultured in DMEM (Invitrogen) media supplemented with 10% FBS, 100 $\mu$g/mL penicillin, 100 $\mu$g/mL streptomycin, 10 mM HEPES, and 6 mM L-glutamine. The HT29$^{\mathrm{mock}}$ and HT29$^{\beta 6\mathrm{AS}}$ cells were cultured in RPMI media (Invitrogen, San Diego, CA) supplemented with 10% FBS and 2.5 $\mu$g/mL puromycin. The SW480$^{\beta 6\mathrm{OE}}$ and SW480$^{\mathrm{mock}}$ cells were cultured in DMEM supplemented with 4.5 g/L glucose, 10% FBS, and 500 $\mu$g/mL Geneticin G418 (Invitrogen). The cells were seeded at $2 \times 10^5$ cells/mL and were grown until ~50% confluence prior to immunofluorescence and PLA experiments. All cells were grown at 37 °C in 5% $CO_2$ (v/v) in biological triplicates.

### Immunofluorescence (IF)

The presence and/or absence of uPAR and $\beta 6$ in all five cell lines were confirmed using IF. When cell cultures reached ~50% confluence, the cells were fixed using 2% paraformaldehyde for 10 min, washed with 0.1 M glycine in PBS, and incubated with blocking solution (9% goat serum, 1% BSA in PBS) for 1 h at room temperature. The cells were then incubated with anti-uPAR R4 (5 $\mu$g/mL) and anti-$\alpha v\beta 6$ 6.4B4 (5 $\mu$g/mL) antibodies for 1 h at 37 °C followed by incubation with Alexa Fluor 488 goat Anti-Mouse IgG (H+L) (Invitrogen) as secondary antibody (4 $\mu$g/mL), for 1 h at 37 °C. Cell nuclei were counter stained with the blue fluorescent DAPI (Invitrogen) nucleic acid stain (300 nM) for 10 min and mounted on glass slides in Gelmount (ProScitech, Australia). The cells were analyzed using a UPLSAPO 40× objective (NA. 0.95) on a fluorescence microscope (BX63, Olympus, Tokyo). All image capture was conducted using a XM10, monochrome cooled CCD camera and CELLSENS dimensions software (Olympus, Tokyo).

### Proximity Ligation Assay (PLA)

The assay was performed according to manufacturer's instructions (Olink Bioscience, Uppsala, Sweden). Briefly, the PLUS oligonucleotide probe was conjugated to anti-uPAR R4 and its isotype control (IgG1), while the MINUS oligonucleotide probe was conjugated to anti-$\alpha v\beta 6$ 6.4B4 and its corresponding isotype control (IgG1). Cells were fixed using 2% paraformaldehyde in PBS and blocked using blocking solution (9% goat serum, 1% BSA in PBS). Oligonucleotide probe conjugated antibodies were introduced to the cells and incubated for 1 h, followed by incubation with the ligation solution for 30 min, followed by amplification solution (contains Cy5 fluorophore) for 100 min. Cells were counter stained with SYBR Green1 stain and mounted. The PLUS and MINUS oligonucleotide conjugated IgG1 mAbs were used as negative controls.

### PLA Imaging

The cells were imaged using an Olympus Fluoview 300 confocal laser scanning system equipped with an inverted microscope (IX70, Olympus Tokyo). A 40× UPLAN APO objective (NA 0.95) was used for analysis of all slides. SYBR Green1 stain was excited using a 488 nm argon laser and the emission signal detected using 510 and 530 nm interference filters. The Cy5 dye was excited using the 633 nm HeNe laser, and the emission signal was detected using a long pass 610 barrier filter. Three sets of images, in the *X*, *Y*, and *Z* dimensions (10 optical slices with a spacing of 0.5 $\mu$m), were captured for each replicate and image analysis performed on the extended XYZ images, using Duolink Image Tool software (Olink Bioscience). The number of protein interaction signals (seen as red spots) per cell was calculated for each image. Aggregated cells were counted manually to avoid miscalculation. A student *t* test was performed to establish the statistical significance of uPAR·$\alpha v\beta 6$ for each cell line.

### uPAR Peptide Array

A cellulose-bound array of 108 spots of 15-mer peptides covering the complete uPAR sequence of 331 amino acids with a 3 amino acid shift was synthesized using SPOT synthesis.[24,25] The uPAR peptide arrays were blocked with 5% skim milk followed by incubation with HRP conjugated recombinant proteins (HRP-RPs) for 4 h. HRP-RPs were prepared by a Lightning-Link HRP conjugation kit (Innova Biosciences) as per the manufacturer's instructions. Unbound HRP-RPs was washed off, and bound HRP-RPs was detected using Super-Signal West Femto Chemiluminescent Substrate (Thermo Scientific). Images were captured using a Fujifilm CS3000 imager in chemiluminescence mode with the intensity adjusted such that the darkest spots were slightly below saturation. The images were then analyzed using MultiGuage software (FujiFilm). A quantitative intensity value for each spot was calculated using the following formula:

$$\mathrm{intensity} = (\mathrm{AU} - \mathrm{BG})/t$$

where "AU" is the measured intensity of each spot, "BG" is the background, and "*t*" is the time of exposure of the imaging. The uPAR peptide array with $\alpha v\beta 6$ was performed in triplicate to confirm reproducibility.

### Bioinformatics Analysis of uPAR Interaction

The known crystal structures (PDB ID: 3BT1) of uPAR, uPA, and Vn complex[26] were analyzed using the ICM bioinformatics software (Internal Coordinate Mechanics).[27] First, the uPAR regions that bound to $\alpha v\beta 6$ on the peptide array were graphically visualized using ICM. These regions were then subjected to manual analysis to determine residues with favorable side-chain orientations. The residues with favorable side-chain orientations were then reanalyzed to determine $\alpha v\beta 6$ residues potentially accessible to the outer surface of uPAR based on hydrophobicity.

## ■ RESULTS AND DISCUSSION

Previous proteomics studies using immunoprecipitation, mass spectrometry, and Western blot analysis, using the ovarian cancer cell line OVCA429,[5] demonstrated that uPAR potentially interacts with other membrane associated proteins, including the $\alpha v\beta 6$ integrin heterodimer. Many of the proteins identified in that study had been previously implicated in either the biology of cancer metastasis, the regulation of plasminogen activation, or as prognostic indicators of poor cancer patient survival (e.g., $\alpha$-enolase, $\alpha v\beta 6$, uPAR). Specifically, uPAR and $\alpha v\beta 6$ have been independently implicated in both cancer biology (e.g., proliferation, TGF$\beta$ activation, cell adhesion, migration, proteolysis, and invasion) and poor epithelial cancer patient prognosis (colorectal, breast, prostate, lung, and ovarian cancer).[7] Coexpression of uPAR and $\alpha v\beta 6$ in the OVCA429 and other cell lines is now well established.[5] Studies using flow cytometry have also independently confirmed the expression of

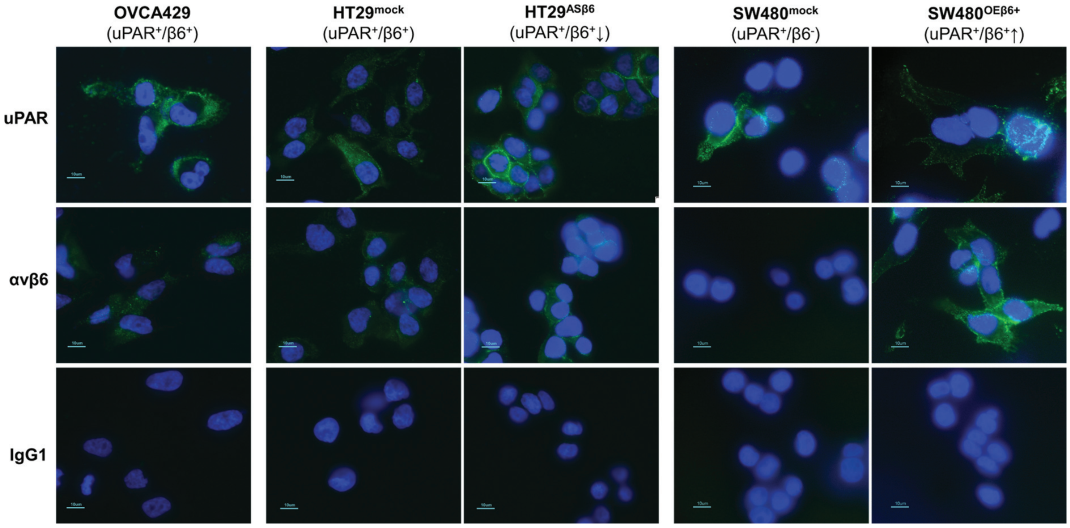


**Figure 1.** A representation of the cell surface expression of uPAR and $\alpha v\beta 6$ for five different cell lines as SW480 $\beta 6$OE, SW480 mock, OVCA-429, HT-29 mock, and HT-29 $\beta 6$AS each expressing varying levels of $\beta 6$. The third row represents the antibody control (IgG1). Nuclei were stained with DAPI, while proteins were detected with a secondary antibody conjugated to Alexa 488.

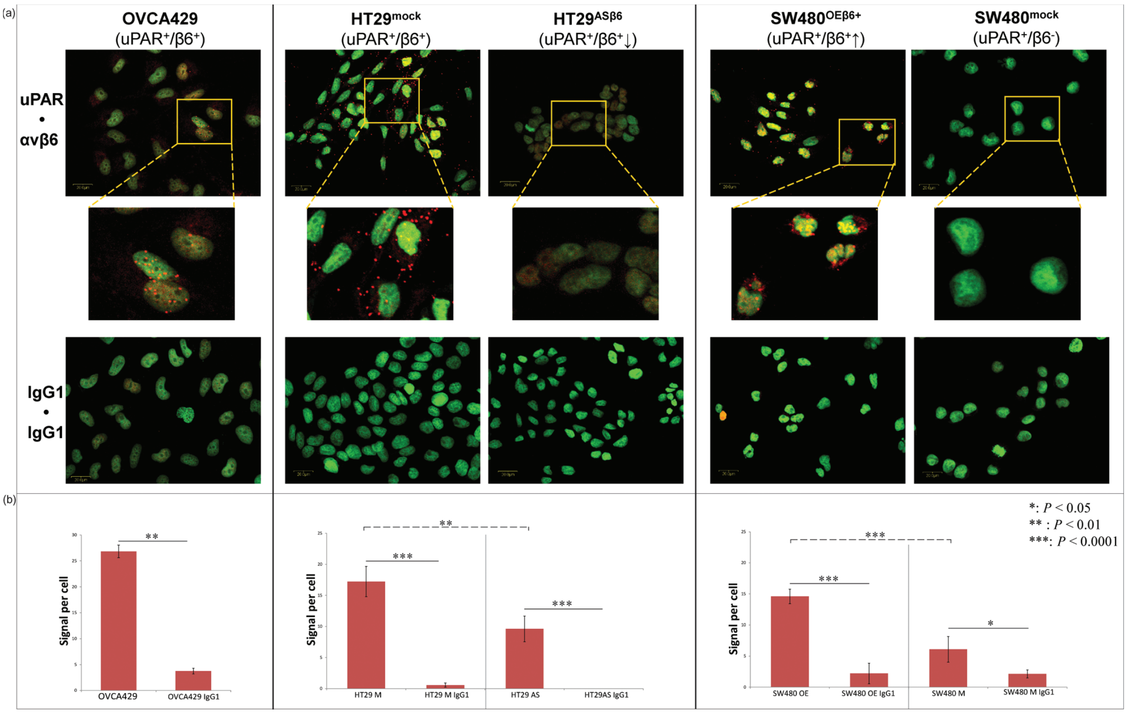


**Figure 2.** Proximity ligation assay images of the cells shown in (A) where the red spots represent the interaction between uPAR·$\alpha v\beta 6$. A signal for the interaction of the uPAR·$\alpha v\beta 6$ corresponding to the level of $\beta 6$ in the cell seems to be observed as compared to the IgG1 isotype control. (B) This observation was quantified by measuring the number of spots per cell. The results showed a significant decrease in interaction when the level of $\beta 6$ was reduced by 35% (in HT-29 $\beta 6$AS cells) ($p < 0.05$). Similarly, a significant increase in interactions was observed when $\beta 6$ was up-regulated in SW480 $\beta 6$OE cells.

both of these antigens on the cell surface.[23,28−30] However, correlations of tumor tissue coexpression and relationships with cancer stage, differentiation status, and patient clinical outcomes (including survival) remain to be explored. The confirmation of a direct uPAR·$\alpha v\beta 6$ interaction would suggest a novel paradigm that potentially explains how and why these membrane proteins share critical aspects of tumor biology and would assist in the development of novel therapeutics to prevent cancer metastasis.[29]

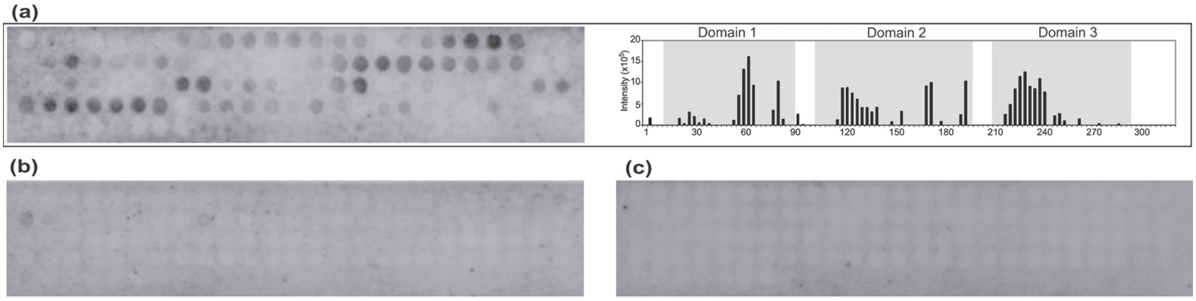


**Figure 3.** (a) uPAR peptide array incubated with $\alpha v\beta 6$ and corresponding intensity plot indicating locations of binding on the three domains of uPAR with the more intense spot (semiquantitatively indicated on the bar chart) indicating a stronger affinity for the heterodimer to the corresponding uPAR peptide. The same peptide array incubated with $\alpha v$ (b) and $\beta 6$ (c) integrins separately, neither of which showed any binding to the array.

The aim of the present study was to functionally validate our previous proteomic studies[5] on IP pull downs of the specific interacting sites of uPAR·$\alpha v\beta 6$ by using two diverse orthogonal biochemical techniques: PLA for in cellulo studies and peptide arrays for in vitro analysis of the specific interacting sites. To validate the uPAR·$\alpha v\beta 6$ interaction, ovarian (OVCA429) and four colon cancer cell lines were employed (HT29$^{\text{mock}}$, HT29 $^{\beta 6\text{AS}}$, SW480$^{\beta 6\text{OE}}$, and SW480$^{\text{mock}}$). The dysregulation of uPAR and $\beta 6$ in these cell lines has been previously demonstrated by various techniques not limited to but including flow cytometry, Western blot, and PET analysis.[29,31−33]

## Immunofluorescence and PLA Confirm the Presence of uPAR·$\alpha v\beta 6$ Interactions

In this study, immunofluorescence (IF) was used to demonstrate the presence of uPAR and $\alpha v\beta 6$ on the cell surface using anti-uPAR R4 and anti-$\alpha v\beta 6$ 6.4B4 mAbs. Consistent with previous studies, these results demonstrated that uPAR was expressed on the cell surface of all cell lines, while $\alpha v\beta 6$ was expressed on SW480$^{\beta 6\text{OE}}$, HT29$^{\text{mock}}$, HT29$^{\beta 6\text{AS}}$, and OVCA429, but was not on SW480$^{\text{mock}}$ (Figure 1). No binding (no fluorescence) was observed with the negative isotype control IgG1 antibody (Figure 1) as control.

Proximity ligation is an emerging technology that has been used to visualize and simultaneously quantify P·P interactions occurring in situ.[34] Proteins in close proximity (30−40 nm) are fluorescently detected using rolling circle amplification of ligatable DNA primers attached to secondary antibodies that bind a pair of epitope-specific monoclonal antibodies.[34,35] In our study, primary antibodies were directed against uPAR and $\alpha v\beta 6$. Expression of integrin $\beta 6$ is restricted to epithelial cells, and it is only known to dimerize with the $\alpha v$ subunit.[36] Therefore, to identify whether interaction with uPAR could be demonstrated quantitatively, we examined other cell lines in which relative expression levels of the $\beta 6$ integrin were modulated. The cell lines used expressed uPAR with varying levels of integrin $\beta 6$ expression. For example, cells that did not express $\beta 6$ (i.e., SW480$^{\text{mock}}$) were compared to those in which integrin $\beta 6$ had been engineered to be overexpressed (SW480$^{\beta 6\text{OE}}$). In addition, cells that endogenously expressed $\beta 6$ (HT29$^{\text{mock}}$) were compared to subclones of the same cell line in which $\beta 6$ expression had been deliberately and stably reduced by ~80% (i.e., HT29$^{\beta 6\text{AS}}$)[29] (Figure 2).

To allow statistical analyses, the assay was performed in biological triplicate for all cell lines, and three images were acquired for each replicate. A significant number of positive spots were observed localized to the cell surface as anticipated (Figure 2). The OVCA429, SW480$^{\beta 6\text{OE}}$, and HT29$^{\text{mock}}$ cell lines showed strong signals for the uPAR·$\alpha v\beta 6$ interaction, whereas the HT29$^{\beta 6\text{AS}}$ cell line showed much weaker signals ($p < 0.05$) (Figure 2a), which is in agreement with the reduced $\beta 6$ expression previously reported.[29] The SW480$^{\text{mock}}$ cell line, where $\beta 6$ is completely absent, showed no apparent uPAR·$\alpha v\beta 6$ PLA signal (Figure 2a). An analysis of the average signal obtained per cell as compared to the corresponding isotype controls demonstrated that the signals obtained from uPAR·$\alpha v\beta 6$ were significantly greater ($p < 0.05$) than the control (Figure 2b).

The results for the OVCA429 cell line were similar to those we had obtained previously.[5] For the colon cancer cell lines, PLA data showed a significant decrease in interaction when the level of $\alpha v\beta 6$ was reduced; concordantly, a significant increase in interaction was observed when $\alpha v\beta 6$ was up-regulated.

In all cases, our PLA results were in good agreement with previous expression data,[29] showing that quantitative uPAR·$\alpha v\beta 6$ PLA signal could be altered simply by decreasing or increasing the expression level of $\beta 6$ present on the cell surface. All isotype controls were negative. However, while collectively these data show close proximity of uPAR and $\beta 6$ indicative of an interaction, the possibility that other "bridging" proteins may be involved in direct interactions with either partner in uPAR·$\alpha v\beta 6$ could not be conclusively excluded. To eliminate this possibility, direct uPAR·$\alpha v\beta 6$ was probed using an orthogonal technique, peptide arrays.

## Peptide Arrays Map Potential Sites of uPAR·$\alpha v\beta 6$ Interaction

Peptide arrays are cost-efficient, accurate, and reliable one-dimensional reconstructions that allow mapping of potential peptidyl binding sites of labeled full length interacting proteins.[37] They have been widely used to analyze large arrays of synthetic peptides on cellulose membranes, facilitating the rapid screening of diverse biomolecule probes.[38] SPOT synthesis[24] was used in this study to generate an array composed of 108 sequential overlapping (3 residues) 15-mer peptides (along the linear uPAR expressed protein sequence) arranged successively on a cellulose membrane. This was used to map the potential binding sites of uPAR and the heterodimeric $\alpha v\beta 6$ integrin, as well as the individual integrin subunits ($\alpha v$ and $\beta 6$). While this method involves a reduction of the three-dimensional uPAR structure into single linear overlapping 15-mer peptides, the method has been used

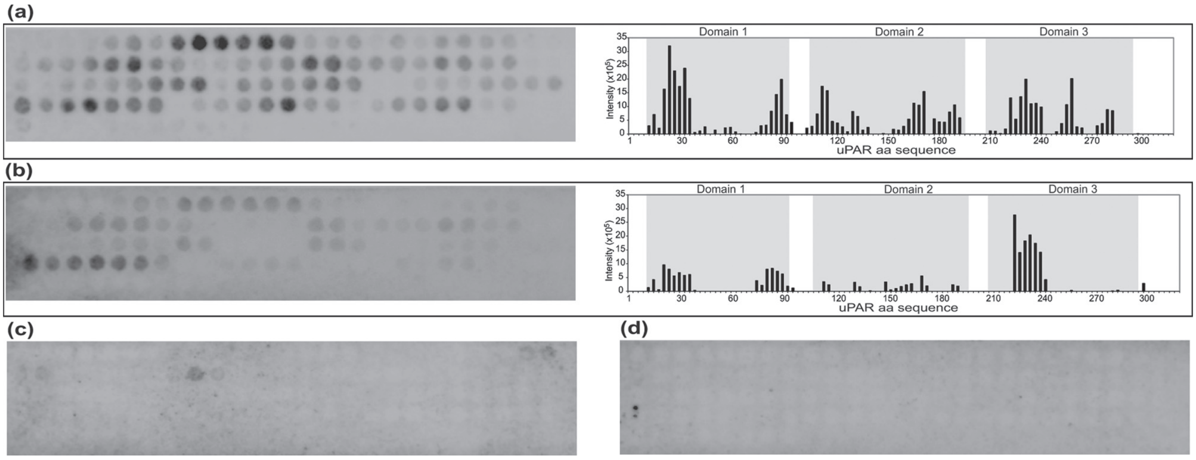


**Figure 4.** (a) uPAR peptide array incubated with uPA and corresponding intensity plot indicating locations of binding on the three domains of uPAR with the more intense spot (semiquantitatively indicated on the bar chart) indicating a stronger affinity for the heterodimer to the corresponding uPAR peptide. The same peptide array incubated with vitronectin, another known binding partner of uPAR, and its corresponding intensity plot (b) and the $\beta$1 (c) and $\beta$3 (d) integrins separately, neither of which again, as monomers, showed any binding to the array.

**Table 1. Potential uPAR and Integrin $\alpha v\beta 6$ Interaction Sites[a]**

| uPAR domain | region identified from peptide array | possible surface residues identified | overlapping residues binding to Vn (uPA) |
|---|---|---|---|
| I | **61** ELVEKSCTHSEKTNRTLS **78** | E61, V63, K65, S70, E71, N74, T76, S78 | S78 (T76) |
| | **82** GLKITSLTEVVCGLD **96** | I85, S87, T89, V91, L95 | I85, S87 (T89) |
| II | **121** GSSDMSCERGRHQSLQCRSPE **141** | M125, R129, R131, H132, S134, Q136, R138 | Q136, R138 |
| | **172**LPGCPGSNGFHNNDTFHF **189** | S178, N184, D185, F187, F189 | none |
| | **193** CNTTKCNEGPILELE **207** | N194, T195, K197, E200, P202, E207, N208 | none |
| III | **229** SEETFLIDCRGPMNQCLVATGTHEPKN **255** | S229, E230, L234, D236, D238, N242, Q243, V246, T248, T250, T254 | none |

[a]Regions binding to integrin $\alpha v\beta 6$ on the peptide array and possible surface residues were identified by manual analysis of the uPAR crystal structure. The last column lists known overlapping binding residues to Vn and uPA (in parentheses). Amino acid residue numbers correspond to full uPAR sequence from UniProt KB (ID: Q03405).

successfully to identify linear specific binding sequences involved in many P·P interactions.[24]

In this study, a GUI (graphical user interface) was developed to semiquantitatively determine the binding affinity of the labeled species (e.g., HRP-labeled $\alpha v\beta 6$) to the uPAR peptide array based on the intensity of positive spots identified (Figure 3a). Overall, our data showed that integrin $\alpha v\beta 6$ binds to peptides emanating from all three uPAR domains (DI, DII, and DIII); in particular, positive binding of labeled-$\alpha v\beta 6$ was located within the following uPAR amino acid sequences: uPAR DI at E61-R75 and G82-D96, uPAR DII at G121-E141, L172-F189, and C193-E207, and uPAR DIII at S229-N255.

In control experiments using identical protein concentrations, the individual integrin protein subunits $\alpha v$ (Figure 3b) or $\beta 6$ (Figure 3c) did not bind to any region of the uPAR peptide array, in contrast to the $\alpha v\beta 6$ dimer.

The peptide array was also used to identify the binding sites of other potential uPAR partners, uPAR's cognate protease ligand uPA and the well-established binding partner Vn. The integrin subunits $\beta 1$ and $\beta 3$ were also examined to determine if they were able to bind as individual integrin subunits in contrast to the data observed for $\beta 6$ (Figure 3C).

These data showed that uPA could bind through domain I, C16-V51, I85-T108; domain II, S112-H150, C169-P210; and domain III, M226-Y258 and I283-V300, (Figure 4a), while Vn was found to bind to domain I, G22-V51, G82-R105; domain II, L116-H150, L172-E207; and domain III, G226-N255 (Figure 4b). As observed for individual subunits $\alpha v$ and $\beta 6$, neither $\beta 1$ nor $\beta 3$ (Figure 4c and d) showed any detectable binding to the uPAR peptide array.

## Structural Mapping of Interacting Sites Reveals Pockets of uPAR·$\alpha v\beta 6$ Interactions

Six potential binding sites were located on the uPAR sequence from the collective peptide array data. These sites were found to be spread across all three domains of uPAR and covered almost 35% of the uPAR sequence. Interestingly, a number of the sequences found to bind to $\alpha v\beta 6$ integrin have previously been implicated in interactions with either Vn and/or uPA (Table 1).[7] To narrow potential docking/binding sites for integrin $\alpha v\beta 6$, an in silico structural analysis of where these six sites were located on the uPAR crystal structure was undertaken and mapped using ICM software (Figure 5a). This was followed by a manual identification of uPAR regions with residues containing favorable side-chain orientations and then investigated for potential residues that could be accessed on the outer surfaces of uPAR (Table 1).

Initial uPAR residue side-chain orientation analysis revealed that approximately 39% of the $\alpha v\beta 6$ interacting uPAR residues identified on peptide arrays possessed side chains found in favorable orientations (i.e., surface accessible). However,

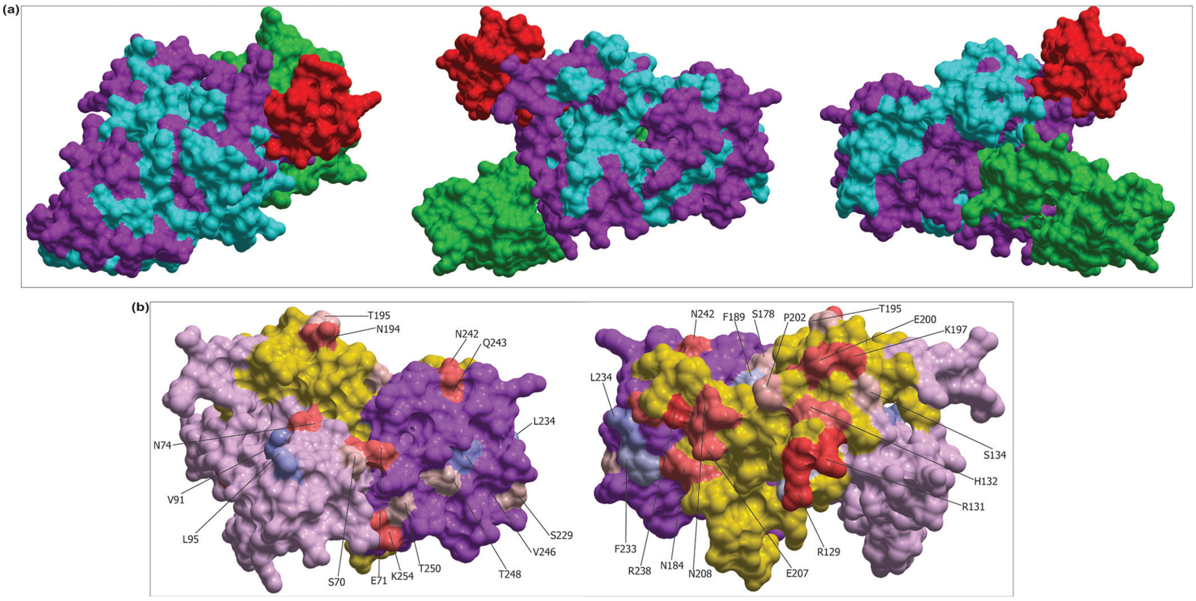


**Figure 5.** (a) The space-filling crystal structure of uPAR (magenta) with red indicating vitronectin, green indicating uPA, and cyan showing regions of uPAR binding to $\alpha v\beta 6$ from uPAR peptide array in three different views. (b) Crystal structure of uPAR only indicating its three domains (light pink, domain I; yellow, domain II; magenta, domain III) overlaid with predicted hydrophobicity labeled in red (residues with H-bond acceptor potential) and blue (residues with H-bond donor potential). The intensity of red and blue shows how strong or weak the H-bond formation potential is, and the numbers correspond to the amino acid sequence of uPAR without the signal peptide. A total of 14 potential residues as sites of binding can be observed on domain II, while 11 can be observed on domain III.

further manual analysis revealed that many of these residues were inaccessible. Only the favorable residues were then subjected to physicochemical (hydrophobicity) analysis (Figure 5). Figure 5b illustrates the hydrophobic nature of the residues identified. It was noted that most of the identified residues had hydrogen (H-) bond acceptor potential (red residues) with some residues having the potential to be H-bond donors (blue residues), while very few residues showed any potential to form H-bonds. Those with acceptor or donor H-bond potentials should prove better binding sites than those with low or no H-bond acceptor potential.

It was clear from this analysis that some residues identified in regions of uPAR domain I (E61 to R75 and G82 to D96) that had been previously suggested to be required for interaction with Vn and/or the receptor's cognate protease ligand uPA[26,39] were buried inside the outer surfaces of uPAR. Residues Q136 and R128, and L172, P173, and H188 in uPAR domain II, which have been previously demonstrated to be required for interaction with Vn and uPA, respectively, were found to be surface accessible.[26,39]

This study revealed that most of the domain II and III residues identified from the arrays could potentially be sites of $\alpha v\beta 6$ integrin interaction. Interestingly, a previous study addressing interactions between integrin $\alpha 5\beta 1$ and uPAR suggested that integrin $\alpha 5\beta 1$ directly interacts with uPAR domain III across the sequence G262-Q270 and the interaction was lost when a single amino acid alanine substitution (S267A) was introduced.[40] Our data suggest that although domains II and III maybe accessible for integrin binding, domain III appears to be a more favorable site, should other ligands be available.

While binding of uPA to its cognate receptor uPAR is a high affinity interaction ($K_d = 4 \times 10^{-10}$ M),[41] significant external regions of uPAR remain available for binding to other potential interacting partners (e.g., Vn and various integrins like $\alpha 3\beta 1$, $\alpha_M\beta 2$, $\alpha v\beta 1$, $\alpha 5\beta 1$, $\alpha v\beta 3$[42]). The uPA and Vn sites indicated from the peptide array showed ~70% overlap with binding sites already published,[26,39] including data obtained from alanine scanning mutagenesis experiments.[9] A detailed structural docking study has been performed to recapitulate and confirm these findings on the interaction of uPAR and $\alpha v\beta 6$.[43]

## ■ IMPLICATIONS AND FUTURE DIRECTIONS

The most likely binding sites for $\alpha v\beta 6$ to uPAR, based on the crystal structure of uPAR (bound to uPA and Vn) coupled with information arising from our peptide array data and a manual analysis of potential binding sites by side-chain orientation and hydrophobicity, appeared to be neighboring adjacent integrin binding sites that were previously identified.[40] An additional advantage of the use of peptide arrays in this study over screening by site directed protein−protein interaction libraries or molecular modeling is that not only are potential binding sites identified, but lead peptide antagonists also determined. These can subsequently be used as tools to address the specific interaction under study.[44] Structural analysis coupled with the previous study on interaction of uPAR with $\alpha 5\beta 1$[40] suggests that uPAR domain III may be a favorable binding site for "all" uPAR-binding integrins. Experiments using blocking peptides against the domain III region of uPAR to determine the precise binding site of uPAR and integrin $\alpha v\beta 6$ are currently ongoing.

For cell motility, invasion, proliferation, and adhesion, it is essential for uPAR to interact with transmembrane proteins for transmission of specific signals across cell membranes to activate appropriate intracellular second messenger systems. Thus, interaction of uPAR with $\alpha v\beta 6$ and other integrins not only couples the proteolytic activation (by binding with uPA) with cell signaling but also localizes the proteolysis to the cell surface.[7] Interactions between uPAR and $\alpha v\beta 6$ could potentially have profound implications on the promotion of cancer cell metastasis by activating a series of specific signaling pathways. For example, uPAR is involved in the Ras-ERK pathway, which is known to directly induce EMT in cells.[7,45] The association of uPAR with integrins like $\alpha 3\beta 1$, $\alpha v\beta 1$, $\alpha 5\beta 1$, $\alpha v\beta 3$ has been studied to varying degrees. It has been shown that uPAR interaction with $\beta 1$ activates both FAK and ERK/MAPK pathways,[40] while interaction with $\beta 3$ activates the Rac pathway.[46] Similarly, studies have shown that disruption of a uPAR and $\alpha v\beta 3$ integrin interaction selectively inhibits Vn-induced cell migration,[9,47] implying that $\alpha v\beta 6$ might also modulate cell migration in some comparable manner.

High expression of $\alpha v\beta 6$ is associated with poor prognosis in many cancer types, including colon cancer.[48] Several studies have implicated $\beta 6$ in cell proliferation, migration, and invasion,[49−51] although the mechanisms by which these processes occur remain unclear. Some reports have suggested involvement of $\alpha v\beta 6$ in MMP pathways as a means by which ECM degradation is facilitated.[16,52] For example, Fyn kinase, which associates with $\alpha v\beta 6$, recruits FAK, thereby activating the Rac/ERK/MAPK pathways, which in turn activate MMP3.[50] There is also evidence showing that $\alpha v\beta 6$ activates transforming growth factor TGF$\beta 1$ by a mechanism involving torsional stress (not proteolysis), which leads to up-regulation of MMP pathways.[53] In addition, a direct interaction between $\alpha v\beta 6$-P-ERK2 has been conclusively established[29] and shown to mediate MMP-9 secretion in colon cancer cells.[29]

It is possible that the pathways activated, seemingly independently by uPAR and $\alpha v\beta 6$, could indeed be activated collectively with proteins found in membranes forming the uPAR·$\alpha v\beta 6$ complex. Indeed, in our initial study several other proteins were identified by proteomics to be binding to uPAR.[5] Targeting $\alpha v\beta 6$ integrin has the additional benefit that it is exclusively expressed in epithelial restricted tumors. It is possible that by therapeutically targeting the uPAR·$\alpha v\beta 6$, the $\alpha v\beta 6$ signaling pathway can be uncoupled from the plasmin activity, potentially leading to a disruption of the pathways involved in EMT resulting in decreased metastasis.

This study provides the detailed groundwork for an analysis of the uPAR·$\alpha v\beta 6$ interaction aimed at using it as a potential novel therapeutic cancer target. Further alternative and complementary techniques could be used to elucidate P·P interactions and to identify significant pathways affected by the interaction. When combined with the approaches taken here, methods like cross-linking mass spectrometry[54] in conjunction with competition studies using peptide arrays and surface plasmon resonance analysis (e.g., BIAcore, Proteon) could be used to analyze the binding kinetics of potential interactants. Indeed, preliminary studies using complementary peptides to block the sites of binding followed by functional assays (migration, proliferation, etc.) on related cell lines have been shown to induce biological and morphological effects (data not shown). The consequences of ablating such interactions can be investigated in mouse models of CRC enabling an in vivo approach.

## AUTHOR INFORMATION

### Corresponding Author

*Phone: +61 2 9850 8211. Fax: +61 2 9812-3600. E-mail: mark.baker@mq.edu.au.

### Author Contributions

$^\nabla$These authors contributed equally.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

We thank Paul H. Weinreb and Sheila M. Violette, from Biogen Idec Inc., Cambridge Center, Cambridge, MA 02142, for kindly providing the 6.4B4 antibody against the integrin $\alpha v\beta 6$. This study was supported with research project grant funding from the NHMRC (#1010303), Cancer Council NSW (RG10-04 and RG08-16), and a Macquarie University MQSN grant and supported through the Australian School of Advanced Medicine (ASAM), Macquarie University, MQ Biofocus and Biomolecular Frontiers Research Centres. Some of the research described herein was facilitated by access to the Australian Proteome Analysis Facility (APAF) and Monash University Antibody Technology Facility (MATF), both established under the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).

## REFERENCES

(1) Nguyen, D. X.; Bos, P. D.; Massague, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **2009**, *9*, 274−84.

(2) Kalluri, R.; Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **2009**, *119*, 1420−8.

(3) Pepper, M. S. Role of the matrix metalloproteinase and plasminogen activator-plasmin systems in angiogenesis. *Arterioscler., Thromb., Vasc. Biol.* **2001**, *21*, 1104−17.

(4) Cox, G.; Steward, W. P.; O'Byrne, K. J. The plasmin cascade and matrix metalloproteinases in non-small cell lung cancer. *Thorax* **1999**, *54*, 169−79.

(5) Saldanha, R. G.; Molloy, M. P.; Bdeir, K.; Cines, D. B.; Song, X.; Uitto, P. M.; Weinreb, P. H.; Violette, S. M.; Baker, M. S. Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. *J. Proteome Res.* **2007**, *6*, 1016−28.

(6) Fagerberg, L.; Jonasson, K.; von Heijne, G.; Uhlen, M.; Berglund, L. Prediction of the human membrane proteome. *Proteomics* **2010**, *10*, 1141−9.

(7) Smith, H. W.; Marshall, C. J. Regulation of cell signalling by uPAR. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 23−36.

(8) Llinas, P.; Le Du, M. H.; Gardsvoll, H.; Dano, K.; Ploug, M.; Gilquin, B.; Stura, E. A.; Menez, A. Crystal structure of the human urokinase plasminogen activator receptor bound to an antagonist peptide. *EMBO J.* **2005**, *24*, 1655−63.

(9) Eden, G.; Archinti, M.; Furlan, F.; Murphy, R.; Degryse, B. The urokinase receptor interactome. *Curr. Pharm. Des.* **2011**, *17*, 1874−89.

(10) Mekkawy, A. H.; Morris, D. L.; Pourgholami, M. H. Urokinase plasminogen activator system as a potential target for cancer therapy. *Future Oncol.* **2009**, *5*, 1487−99.

(11) Seetoo, D. Q.; Crowe, P. J.; Russell, P. J.; Yang, J. L. Quantitative expression of protein markers of plasminogen activation system in prognosis of colorectal cancer. *J. Surg. Oncol.* **2003**, *82*, 184−93.

(12) Rabbani, S. A.; Mazar, A. P. The role of the plasminogen activation system in angiogenesis and metastasis. *Surg. Oncol. Clin. N. Am.* **2001**, *10*, 393−415.

(13) Hynes, R. O. Integrins: versatility, modulation, and signaling in cell adhesion. *Cell* **1992**, *69*, 11−25.

(14) Giancotti, F. G.; Ruoslahti, E. Integrin signaling. *Science* **1999**, *285*, 1028−32.

(15) Liu, S.; Liang, B.; Gao, H.; Zhang, F.; Wang, B.; Dong, X.; Niu, J. Integrin alphavbeta6 as a novel marker for diagnosis and metastatic potential of thyroid carcinoma. *Head Neck Oncol.* **2013**, *5*, 7.

(16) Bandyopadhyay, A.; Raghavan, S. Defining the role of integrin alphavbeta6 in cancer. *Curr. Drug Targets* **2009**, *10*, 645−52.

(17) Bates, R. C. The alphaVbeta6 integrin as a novel molecular target for colorectal cancer. *Future Oncol.* **2005**, *1*, 821−8.

(18) Annes, J. P.; Munger, J. S.; Rifkin, D. B. Making sense of latent TGFbeta activation. *J. Cell Sci.* **2003**, *116*, 217−24.

(19) Gu, X.; Niu, J.; Dorahy, D. J.; Scott, R.; Agrez, M. V. Integrin alpha(v)beta6-associated ERK2 mediates MMP-9 secretion in colon cancer cells. *Br. J. Cancer* **2002**, *87*, 348−51.

(20) Weinreb, P. H.; Simon, K. J.; Rayhorn, P.; Yang, W. J.; Leone, D. R.; Dolinski, B. M.; Pearse, B. R.; Yokota, Y.; Kawakatsu, H.; Atakilit, A.; Sheppard, D.; Violette, S. M. Function-blocking integrin alpha(v)-beta(6) monoclonal antibodies - Distinct ligand-mimetic and non-ligand-mimetic classes. *J. Biol. Chem.* **2004**, *279*, 17875−17887.

(21) Tsao, S. W.; Mok, S. C.; Fey, E. G.; Fletcher, J. A.; Wan, T. S.; Chew, E. C.; Muto, M. G.; Knapp, R. C.; Berkowitz, R. S. Characterization of human ovarian surface epithelial cells immortalized by human papilloma viral oncogenes (HPV-E6E7 ORFs). *Exp. Cell Res.* **1995**, *218*, 499−507.

(22) Agrez, M.; Chen, A.; Cone, R. I.; Pytela, R.; Sheppard, D. The alpha v beta 6 integrin promotes proliferation of colon carcinoma cells through a unique region of the beta 6 cytoplasmic domain. *J. Cell Biol.* **1994**, *127*, 547−56.

(23) Weinacker, A.; Chen, A.; Agrez, M.; Cone, R. I.; Nishimura, S.; Wayner, E.; Pytela, R.; Sheppard, D. Role of the integrin alpha v beta 6 in cell attachment to fibronectin. Heterologous expression of intact and secreted forms of the receptor. *J. Biol. Chem.* **1994**, *269*, 6940−8.

(24) Frank, R. The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports–principles and applications. *J. Immunol. methods* **2002**, *267*, 13−26.

(25) Frank, R. Spot-Synthesis - an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* **1992**, *48*, 9217−9232.

(26) Huai, Q.; Zhou, A.; Lin, L.; Mazar, A. P.; Parry, G. C.; Callahan, J.; Shaw, D. E.; Furie, B.; Furie, B. C.; Huang, M. Crystal structures of two human vitronectin, urokinase and urokinase receptor complexes. *Nat. Struct. Mol. Biol.* **2008**, *15*, 422−3.

(27) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM−A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(28) Li, Y.; Wood, N.; Yellowlees, D.; Donnelly, P. K. Cell surface expression of urokinase receptor in normal mammary epithelial cells and breast cancer cell lines. *Anticancer Res.* **1999**, *19*, 1223−8.

(29) Ahmed, N.; Niu, J.; Dorahy, D. J.; Gu, X.; Andrews, S.; Meldrum, C. J.; Scott, R. J.; Baker, M. S.; Macreadie, I. G.; Agrez, M. V. Direct integrin alphavbeta6-ERK binding: implications for tumour growth. *Oncogene* **2002**, *21*, 1370−80.

(30) Niu, J.; Gu, X.; Ahmed, N.; Andrews, S.; Turton, J.; Bates, R.; Agrez, M. The alphaVbeta6 integrin regulates its own expression with cell crowding: implications for tumour progression. *Int. J. Cancer* **2001**, *92*, 40−8.

(31) Moreau, M.; Mourah, S.; Dosquet, C. beta-Catenin and NF-kappaB cooperate to regulate the uPA/uPAR system in cancer cells. *Int. J. Cancer* **2011**, *128*, 1280−92.

(32) Ronne, E.; Behrendt, N.; Ploug, M.; Nielsen, H. J.; Wollisch, E.; Weidle, U.; Dano, K.; Hoyer-Hansen, G. Quantitation of the receptor for urokinase plasminogen activator by enzyme-linked immunosorbent assay. *J. Immunol. Methods* **1994**, *167*, 91−101.

(33) Persson, M.; Madsen, J.; Ostergaard, S.; Jensen, M. M.; Jorgensen, J. T.; Juhl, K.; Lehmann, C.; Ploug, M.; Kjaer, A. Quantitative PET of human urokinase-type plasminogen activator receptor with 64Cu-DOTA-AE105: implications for visualizing cancer invasion. *J. Nucl. Med.* **2012**, *53*, 138−45.

(34) Weibrecht, I.; Leuchowius, K. J.; Clausson, C. M.; Conze, T.; Jarvius, M.; Howell, W. M.; Kamali-Moghaddam, M.; Soderberg, O. Proximity ligation assays: a recent addition to the proteomics toolbox. *Expert Rev. Proteomics* **2010**, *7*, 401−9.

(35) Thymiakou, E.; Episkopou, V. Detection of signaling effector-complexes downstream of bmp4 using PLA, a proximity ligation assay. *J. Visualized Exp.* **2011**.

(36) Breuss, J. M.; Gillett, N.; Lu, L.; Sheppard, D.; Pytela, R. Restricted distribution of integrin beta 6 mRNA in primate epithelial tissues. *J. Histochem. Cytochem.* **1993**, *41*, 1521−7.

(37) Li, S. S.; Wu, C. Using peptide array to identify binding motifs and interaction networks for modular domains. *Methods Mol. Biol.* **2009**, *570*, 67−76.

(38) Maier, R. H.; Maier, C. J.; Rid, R.; Hintner, H.; Bauer, J. W.; Onder, K. Epitope mapping of antibodies using a cell array-based polypeptide library. *J. Biomol. Screening* **2010**, *15*, 418−26.

(39) Barinka, C.; Parry, G.; Callahan, J.; Shaw, D. E.; Kuo, A.; Bdeir, K.; Cines, D. B.; Mazar, A.; Lubkowski, J. Structural basis of interaction between urokinase-type plasminogen activator and its receptor. *J. Mol. Biol.* **2006**, *363*, 482−95.

(40) Chaurasia, P.; Aguirre-Ghiso, J. A.; Liang, O. D.; Gardsvoll, H.; Ploug, M.; Ossowski, L. A region in urokinase plasminogen receptor domain III controlling a functional association with alpha5beta1 integrin and tumor growth. *J. Biol. Chem.* **2006**, *281*, 14852−63.

(41) Vassalli, J. D.; Baccino, D.; Belin, D. A cellular binding site for the Mr 55,000 form of the human plasminogen activator, urokinase. *J. Cell Biol.* **1985**, *100*, 86−92.

(42) Blasi, F.; Carmeliet, P. uPAR: a versatile signalling orchestrator. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 932−43.

(43) Sowmya, G.; Khan, J. M.; Anand, S.; Ahn, S. B.; Baker, M. S.; Ranganathan, S. A site for direct integrin alphavbeta6.uPAR interaction from structural modelling and docking. *J. Struct. Biol.* **2014**, *185*, 327−35.

(44) Hruby, V. J. Designing peptide receptor agonists and antagonists. *Nat. Rev. Drug Discovery* **2002**, *1*, 847−58.

(45) Jo, M.; Eastman, B. M.; Webb, D. L.; Stoletov, K.; Klemke, R.; Gonias, S. L. Cell signaling by urokinase-type plasminogen activator receptor induces stem cell-like properties in breast cancer cells. *Cancer Res.* **2010**, *70*, 8948−58.

(46) Smith, H. W.; Marra, P.; Marshall, C. J. uPAR promotes formation of the p130Cas-Crk complex to activate Rac through DOCK180. *J. Cell Biol.* **2008**, *182*, 777−90.

(47) Degryse, B.; Orlando, S.; Resnati, M.; Rabbani, S. A.; Blasi, F. Urokinase/urokinase receptor and vitronectin/alpha(v)beta(3) integrin induce chemotaxis and cytoskeleton reorganization through different signaling pathways. *Oncogene* **2001**, *20*, 2032−43.

(48) Bates, R. C. Colorectal cancer progression: integrin alphavbeta6 and the epithelial-mesenchymal transition (EMT). *Cell Cycle* **2005**, *4*, 1350−2.

(49) Ramos, D. M.; But, M.; Regezi, J.; Schmidt, B. L.; Atakilit, A.; Dang, D.; Ellis, D.; Jordan, R.; Li, X. Expression of integrin beta 6 enhances invasive behavior in oral squamous cell carcinoma. *Matrix Biol.* **2002**, *21*, 297−307.

(50) Li, X.; Yang, Y.; Hu, Y.; Dang, D.; Regezi, J.; Schmidt, B. L.; Atakilit, A.; Chen, B.; Ellis, D.; Ramos, D. M. Alphavbeta6-Fyn signaling promotes oral cancer progression. *J. Biol. Chem.* **2003**, *278*, 41646−53.

(51) Ramos, D. M.; Dang, D.; Sadler, S. The role of the integrin alpha v beta6 in regulating the epithelial to mesenchymal transition in oral cancer. *Anticancer Res.* **2009**, *29*, 125−30.

(52) Morgan, M. R.; Thomas, G. J.; Russell, A.; Hart, I. R.; Marshall, J. F. The integrin cytoplasmic-tail motif EKQKVDLSTDC is sufficient to promote tumor cell invasion mediated by matrix metalloproteinase (MMP)-2 or MMP-9. *J. Biol. Chem.* **2004**, *279*, 26533−9.

(53) Xu, J.; Lamouille, S.; Derynck, R. TGF-beta-induced epithelial to mesenchymal transition. *Cell Res.* **2009**, *19*, 156−72.

(54) Tang, X.; Bruce, J. E. A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. *Mol. BioSyst.* **2010**, *6*, 939−47.