Technische Universität München

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Fachgebiet für Bioinformatik

Classification of protein protein interactions

Florian Goebels

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. A. Tellier
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. D. Frischmann
	2. Univ.-Prof. Dr. R. Zimmer
	(Ludwig-Maximilian-Universität München)

Die Dissertation wurde am 13.08.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 20.10.2014 angenommen.

Abstract

In recent years a wide range of high-throughput methods have been developed for detecting protein interactions, but they only measure whether or not two proteins interact and give no insight into the interaction's characteristics. In this PhD thesis we focused on two common interaction types. We distinguished protein interactions into obligate or non-obligate whether or not the protomers can exist independently. Furthermore, a protein interaction can be simultaneously possible (SP) or mutually exclusive (ME) based on the binding site specificity. There are several reported methods for distinguishing protein interaction types using the three-dimensional structures of protein complexes. In this thesis, we report PiType, a novel technique for classifying protein interactions into obligate/non-obligate, and into SP/ME based exclusively on sequence and network information. Contrary to structure based classifiers, PiType is suitable for large-scale classification of interaction data. Moreover, PiType achieves an auROC of at least 80% and a F-measure close to 80% in cross-fold validation, which is comparable to the performance of structure-based classifiers. We found that the proteins which take part in non-obligate interactions have a higher degree of structural disorder, more short linear motifs, and lower functional similarity compared to obligate interaction partners. As for SP and ME interactions, we observed significant differences in network topology.

In our followup work, we created PiType 2.0, an updated version of PiType. Major enhancements include the integration of the STRING database, which greatly increases the number of classifiable interactions from a much broader range of species. Furthermore, we improved feature calculation methods, and greatly reduced PiType's computational run time. Finally we created a web service for fast and easy use of the PiType 2.0 pipeline, which is freely available for use, and can be downloaded as a self-installing package from http://webclu.bio.wzw.tum.de/PiType/.

Zusammenfassung

Viele der gängigen Hochdurchsatz-Screening Methoden können nur Messen ob zwei Proteine interagieren, liefern jedoch keinen tieferen Einblick in den Interaktionstyp. In dieser Doktorarbeit beschäftigen wir uns mit zwei üblichen Interaktionstypen: Obligaten und nicht obligaten Protein-Protein Interaktionen, definiert dadurch, ob die Proteine unabhängig voneinander existieren können. Außerdem kann ein Protein entweder mit mehreren Partnern gleichzeitig (SP) oder nur mit einem Partner auf einmal (ME) interagieren, abhängig von der Bindestelle. In mehreren Publikationen werden Methoden beschrieben, die in der Lage sind anhand struktureller Informationen Protein Interaktionstypen zu erkennen. In dieser Arbeit beschreiben wir mit PiType eine neue Methode zur strukturunabhängigen Klassifikation von Protein-Protein Interaktionen der oben genannten Typen. Im Gegensatz zu bisherigen Methoden ist PiType in der Lage Protein-Interaktionsnetzwerke im Hochdurchsatz zu klassifizieren. Zusätzlich erreicht PiType einen auROC Wert von mindestens 80% und einen F-Wert von knapp 80%. Dies ist mit strukturbasierten Klassifikationsmethoden vergleichbar. Wir haben gezeigt, dass Proteine in nicht obligaten Interaktionen ein höheres Maß an intrinsischen, ungeordneten Regionen besitzen und mehr kurze, lineare Motive in ihrer Proteinsequenz besitzen. Des Weiteren sind Proteine, die nicht obligat miteinander Interagieren, funktionell von einander verschiedener, als solche die obligat miteinander interagieren. Bei SP/ME Protein Interaktionen haben wir festgestellt, dass sie sich stark in ihrer lokalen Netzwerktopologie unterscheiden.

Darauf aufbauend haben wir einen aktualisierte Version von PiType entwickelt, nämlich PiType 2.0. Die Integration der STRING Datenbank, wodurch die Anzahl der klassifizierbaren Interaktionen dramatisch erhöht wurden, war die zentrale Verbesserung. Wir haben zudem die Rechenzeit für Interaktionsmerkmale deutlich reduziert. Abschließend haben wir den PiType Web Server erstellt, welcher öffentlich unter der Webadresse http://webclu.bio.wzw.tum.de/PiType/ zu finden ist.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 History of protein interactions

In the year 1828 the Dutch chemist Gerardus Johannes Mulder conducted elementary analysis of several proteins and determined that they all share the same empirical number [25, 167], and the Swedish chemist Jöns Jacob Berzelius proposed to name those types molecules as proteins, which is derived from the Greek word πρωτειος (proteios), meaning "of primary importance" [201]. However, only much later, in the year 1876, the term "enzyme" for such biological macromolecules was first described in a study in which the protein trypsin was isolated [67, 129]. In 1906 the first regulatory protein protein interaction (PPI) was reported and analyzed in a publication that describes the inhibition of trypsin, as well as it's quantitative kinetics [99]. Only later in 1913, L. Michaelis, and Miss Maud L. Menten proposed the Michaelis—Menten kinetic, which describes the kinetics of an enzymatic reaction [148]. However, it took until the late 1940s to fully understand the underlying protein interactions between Actin, Myosin, ATP, and ATPase, which are of utmost importance for physical motion [168]. In the 60s, protein interaction which regulates metabolism were identified, and protein interactions were recognized as key players for the dynamic adaption of enzyme activity according to metabolic requirements of the organism [4, 42, 154]. In the following decade the first signalling cascades were discovered [191, 123, 180]. At the beginning of the 90s the importance of protein interaction was well recognized, that became the essential motivation to create high-throughput methods for measuring possible protein protein interaction. This resulted in the development of the yeast two hybrid method [221], which was one of the first high-throughput method for measuring protein interactions. However, this was not the last method to appear [193], and thus we observed an huge increase in measured protein interactions (Figure 1.1), and several complete protein screenings for model organisms [222]. Due to the significant relevance of protein interactions, it has become common that one of the first things to investigate of an unknown protein is its interaction partners. In this thesis we take this question the

next logical level, and investigate how does a protein interact with its partners.

1.2 Direct physical protein interaction types

Proteins are central macromolecules for almost all biological functions, but rarely act by themselves. Most of the molecular processes rely on molecular machines (for an example see Figure 1.2), which are made up from a large number of proteins which bind to each other via direct physical protein protein interactions. Protein protein interactions in its essence are mediated by protein interfaces. Protein interfaces (or binding sites) are certain patches on each of the two proteins' surface, which enables the interaction between two proteins. Most of the protein interfaces are on the protein surface, since exposure is a central requirement for the interaction to form [139]. Due to their importance in mediating protein protein interactions considerable amount of research has been done analysing protein interfaces. Initial analysis of 3D protein structures revealed that protein interfaces are composed of completely buried cores, which are surrounded by partially accessible rims [18, 32]. Furthermore, the average patch size of an interface lies between 1600 and 400 $Å^2$ [40]. Analysis of protein interface sequences revealed that certain amino acid types are more frequent on protein interfaces, and that there are differences in amino acid composition between the core and the rim of the interface [18, 32, 114, 88]. It has been observed that the majority of the binding affinity of the interface is caused by independent small, and highly packed regions, which are called "hot spots" [47]. Hot spots are highly conserved [139], and the underlying kinetics of the hot spots interactions has been well investigated [127].

Almost all of the cellular processes require that proteins specifically recognise a multitude of different interaction partners. Thus, we observe a vast diversity in protein interactions, however all protein interfaces share several common properties, and can be classified into surprisingly few interaction types. A physical protein interaction can be classified into different interaction types depending on many factors [159, 113]. The most basic differentiation is based on the protein complex composition; a complex consisting of only identical proteins is considered to be a homo–oligomers, conversely a complex made of different proteins is defined as a hetero–oligomers. Protein interactions can be differentiated into obligate or non-obligate depending on the lifetime of the interaction. Based on their binding affinity an interaction can be considered to be permanent or transient. The specitivity of the involved protein interfaces determine if the interaction is simultaneously possible or mutually exclusive. Depending on the situation and cellular process certain interaction types are involved, and required. Hence, it is of vital importance to know, characterise, and understand protein protein interaction types and the effects it hast to certain biological processes.

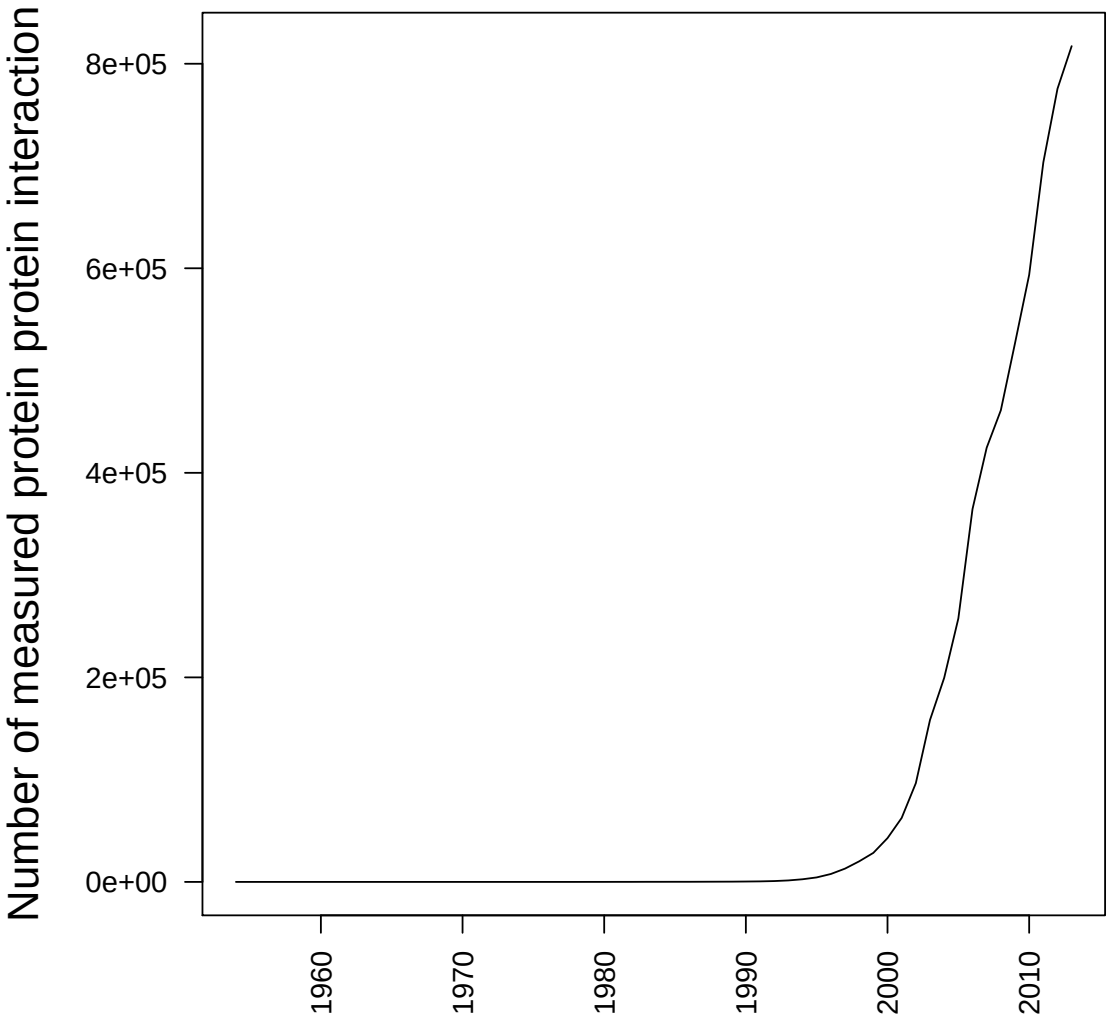


Figure 1.1: Number of stored protein protein interactions (including multiple measurements of the same interaction) in the iRefIndex database [177] for various years.

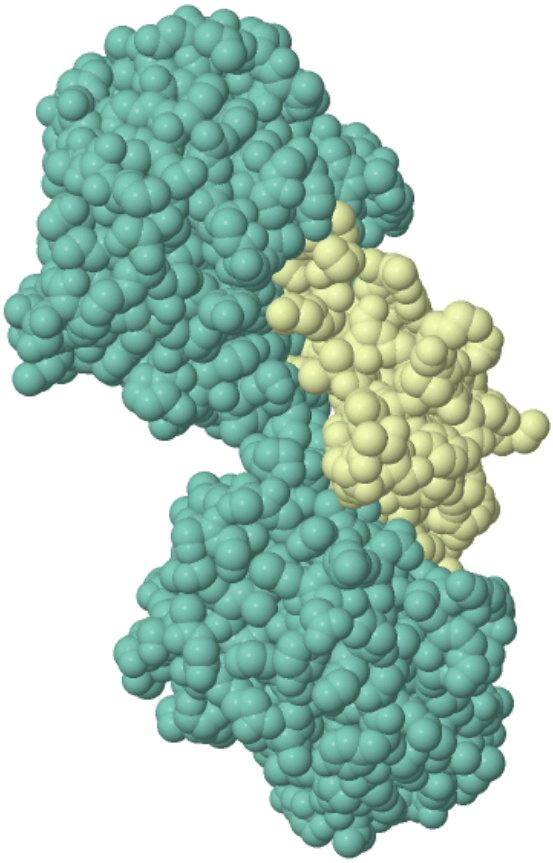


Figure 1.2: Protien interaction between trypsin and its inhibitor (Image was taken from the PDB database [13] with the PDB ID: 1TIM, and was visualized using jsmol [95]).

1.2.1 Homo–oligomers *vs.* hetero–oligomers:

A protein complex is considered to be a homo–oligomers when the complex is composed of only identical interacting sub–units (i.e. proteins), else it is classified as a hetero–oligomers, and thus composed of several different proteins. Homo–oligomers protein complexes tend to form very stable permanent protein structures. This is especially true in cases of homo–dimer (two protein complexes) [113], for example the cytochrome c' homo–dimer [64]. Due to the symmetric nature of homo–oligomers, it is common that they they serve as a scaffold for macromolecules to bind on. One example would be the bacterial GroEL chaperonin protein which can form a zylindric GroEL Homo-Heptamer (seven proteins), that in turn can bind seven GroES proteins on one side of its end [19]. In this work we only consider hetero–dimer protein complexes for classification, since many of the used features exploit the differences between the two interactors (disorderedness, ELM, and functional similarity see section 2.5) and thus can not be applied to homo–dimers.

1.2.2 Obligate *vs.* non-obligate

One important distinction can be made between obligate and non-obligate interactions, dependent on whether or not the protomers can exist independently from each other [159, 166]. The interfaces of non-obligate interactions tend to be smaller, less tightly packed, more polar, less conserved, and overall more similar to normal protein surfaces in terms of amino acid composition than those of obligate interactions [241]. Protein complexes can also be subdivided into two classes based on their binding affinity and lifetime (Figure 1.3). Constituents of permanent interactions, such as enzyme-inhibitors or antibody-antigen complexes, are only found in bound state while transient interactions, usually involved in intracellular signaling, are short-lived and readily associate and dissociate [159]. Interaction sites of transient protein complexes have the tendency to be disordered and their binding specificity is often determined by short linear amino acid motifs (ELM) [166, 51]. Obligate interactions are usually permanent [159] whereas non-obligate interactions are mostly transient [109]. Several machine learning methods have been proposed to automatically classify protein complexes with known three-dimensional structure into various types based on physical, chemical, geometrical, and evolutionary properties of protein recognition sites [17, 241, 164, 136, 183, 141, 23, 161, 151]. For example, Mintseris and Weng achieved an accuracy of 91% in separating transient from permanent complexes using atomic contact vectors to describe the properties of interaction interfaces [151]. Likewise, the NOXclass classifier developed by Zhu et al [241] distinguishes obligate from non-obligate interactions with an accuracy of 91.8% by considering the interface area, amino acid composition, shape complementarity, and evolutionary conservation.

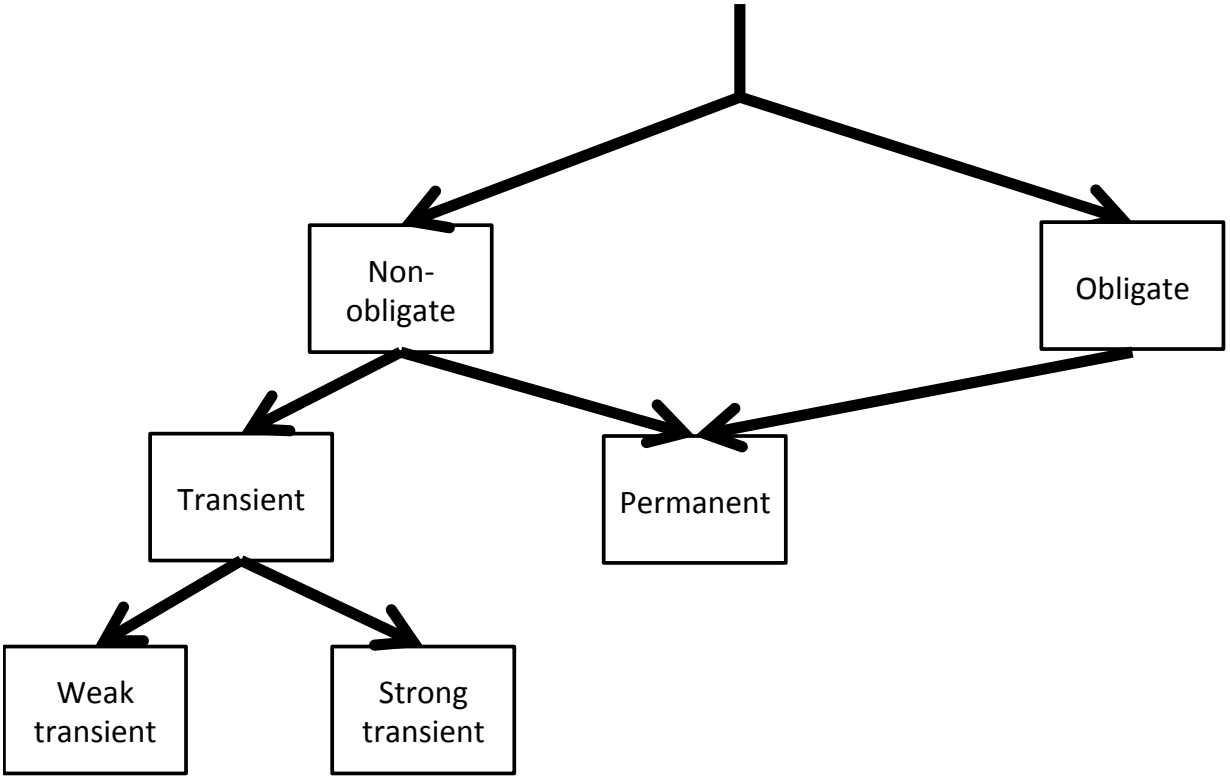


Figure 1.3: Different types of protein interactions. By definition all obligate interactions are permanent; however not all permanent interactions are obligate. Both obligate and permanent interactions tend to have high binding affinities. Non-obligate, and transient interactions tend to have lower binding affinities. The strong transient category includes protein interactions that shift from an unbound/weakly bound state to a strongly bound state, which is generally triggered by an effector molecule.

1.2.3 Permanent *vs.* transient

As mentioned above non-obligate interactions can be further divided into permanent or transient interactions according to the binding affinity (Figure 1.3). However, in this section we will also explain the subdivision into weak and strong transient protein interactions. Weak transient interactions are characterised by having a low binding affinity and very short lifetimes (i.e. several seconds), whereas, Strong transient interactions can change their tertiary structure when triggered (e.g. by ligand binding). In turn, this actives the protein and enables binding. Considerable amount of research has been done investigating transient protein interactions [166]. Similar to obligate, and non-obligate interactions, the structure of transient interfaces tend to be smaller than permanent interfaces, and their amino acid composition is similar to the protein surface, but interfaces are slightly richer in neutral polar groups [160, 2, 48, 114, 40, 152]. Transient interfaces mostly consist of a central core which is completely buried during the interactions, and transient interaction

interfaces tend to be enriched in water. Furthermore, it has been shown that strong transient interactions compared to weak transient interactions have smaller, planar, and more hydrophobic interfaces, and they undergo larger conformational changes [159]. Residues in weak transient interfaces tend to be less stable, but more conserved [160]. Based on those special characteristic features a number of structure dependent methods for permanent/transient classification have been developed [137]. Moreover, a recent found correlation between stability of the unbound residue and transient protein interactions opens new possibilities for the prediction of interfaces from unbound protein structures [20, 53]. It has been shown that intrinsic disorder is strongly associated with transient protein interactions, since disorder-to-order transitions upon binding is often accompanied by a decrease in conformational entropy which in turn causes a low binding affinity [194]. Moreover, disorder was shown to occur frequently in proteins which are associated with cellular processes enriched in transient interactions. This is further emphasised by the fact that long disordered regions can be found on around 60% of all signalling proteins [106]. Signalling pathways are also highly enriched in ELMs [198]. Interactions which are mediated by an ELM normally have very small interfaces [198] and weak binding affinities [49] Because of these properties almost all of the ELM mediated interactions are transient.

1.2.4 Simultaneously possible *vs.* mutually exclusive

Protein interactions can also be classified into two types based on their timing and the spatial distribution of binding sites on the protein surface. Products of co-expressed genes [93] may form stable complexes and interact with each other simultaneously, which is only possible when a network hub ("party hub") possesses a unique binding site for each interaction partner [125]. Alternatively, hub proteins that are not co-expressed with their interaction partners are believed to bind their partners individually at different times (or in different cellular locations) via the same interface ("date hubs") [125]. Following Kim et al. [125] we refer to the interactions of the first and the second type as simultaneously possible (SP) and mutually exclusive (ME), respectively (Figure 1.4). However, to accurate correct prediction of SP/EM interactions one had to analyse pairs of protein protein interactions, but this would be out of scope of this work due to limiting data. In this work we consider an interaction to be ME when at least one of the involved protein binding sites is shared by more than one interaction partner, otherwise the interaction is SP (Figure 1.4). SP and ME interactions and the corresponding binding interfaces can be directly studied by overlaying high-quality protein interaction data with known three-dimensional structures of protein complexes. Analyses of such a structurally resolved interaction network (SIN) together with gene expression patterns revealed distinctly different cellular roles of party and date hubs, with the former corresponding to stable network modules and the latter connecting modules with each other. Date hubs show much lower average degree and are more often encoded by essential genes than party hubs. Similarly, proteins involved in SP interactions (and hence co-expressed) tend to be more functionally similar than those involved in ME interactions, which led to the suggestion that ME interactions are mostly

transient [125] while SP interactions are preferentially obligate [165].

1.3 Genetic interactions/Functional associations

Proteins do not only undergo direct physical interactions, but also can interact indirectly with each another via a functional interplay such as: catalyzation of subsequent events in a pathway, regulation of each other, or they are members of a larger protein complex without being in direct contact [70]. Those kind of interactions are called functional associations (or genetic interactions) [105, 60], and together with direct physical interactions they form the larger superset of "functional protein linkages" [70]. Generally, indirect interactions are not stored in PPI databases, however databases such as BioGrid [33], KEGG [119], Reactome [115], and the famous STRING [70] database store functional associations. In those databases most of the indirect interactions were measured either by synthetic lethality [118] (see Table 1.1), or by co-expression.

Synthetic lethality

Consider two genes a, and b: a mutation in gene a has no effect, also a mutation in gene b has no effect; however, a mutation in both genes a, and b causes a sever phenotype in the host. In cases where the mutations are lethal to the host this is called synthetic lethality [118], and if the mutations cause an unexpected, or sever phenotype (e.g. change in growth rate [61, 172]), it is called a synergistic interaction [142]. Synthetic lethality hints that transcripts of the two genes are physically interacting, share a pathway, or a function in the cell: for example synthetic lethality was used to determine the function of the yeast YLL049W gene, which belongs to the dynein–dynactin pathway, and bridges the mitotic exit network and the Cdc14 early anaphase release pathway [28]. In general synthetic lethality is conduced in a large-scale manner [52], and the preferred method is called synthetic genetic array analysis (SGA) [208]. SGA enables large scale analysis of genetic interactions via crossing a mutated (deletion) query gene of interest against a library of possible gene deletion mutants to make an array of double mutants, which can then be investigated for specific phenotypes [208]. This method was used to create the first genetic interaction map of a cell, and it tested around 4000 possible genetic interactions between 1000 different candidate genes [208]. Consequently, this method was used to create complete genetic interactions map for several model organisms such as yeast [208], E. coli [211], and C. elegans [134]. It goes so far that the current version of the Biogrid database [33] stores a total of 278062 genetic interactions, which is around 40% of all interactions stored in the BioGrid database.

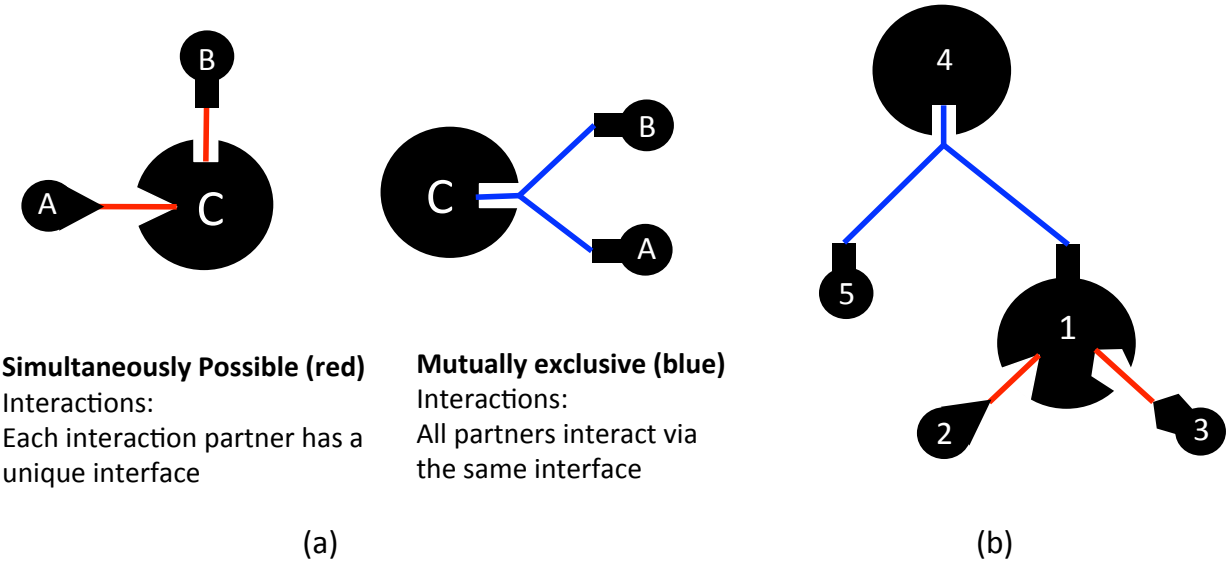


Figure 1.4: Simultaneously possible *vs.* mutually exclusive protein interactions. a) Protein C has two unique binding interface for each of its interaction partners, and thus C can interact simultaneously with both proteins: A and B. In the other case protein C has one binding interface which is shared between C's interaction partners A, and B, and thus only one of them can bind at the same time, which means that both of them are interacting mutually exclusively with C. b) Example network illustrating the interpretation of multiple ME/SP interactions. Consider five proteins 1, 2, 3, 4, and 5. Protein 1 has a different interface for each of its interaction partners (i.e. 2, 3, and 4) and thus can bind all of them at the same time (SP). Protein 4 interacts with its partners 1 and 5 via the same interface and hence can only bind one of them at the same time (ME). While both interactions 1-2 and 1-4 can occur at the same time, the interaction between the proteins 1 and 4 is classified as ME, since the latter protein has only one interface, which is shared between proteins 1 and 5. In conclusion, an interaction between two proteins is considered ME when at least one of the involved protein interfaces is shared by more than one interaction partner, otherwise the interaction is SP.

Co-expression

The second common method for measuring genetic interactions is via analysing the expression profiles of several proteins to distinguish whether or not they are correlated. The basic idea behind co-expression is that the expression profile of genes coding for protein complex subunits should be correlated with the stoichiometric composition of the protein complex [110]. In other words, if a protein complex which consists of three subunits is required by the cell, we expect an increase in expression levels for all three genes. Consequently, if the cell no longer needs the protein complex, the expression levels for all three subunits should decrease. There are several methods for measuring gene expression (either by microarrays [188] or next generation sequencing [143]), and also several methods for calculating gene co-expression [179]. Common methods for measuring an expression profile's similarity involve calculating a correlation coefficient using relative, or normalised expression levels of the gene/proteins in question [210, 126, 239]. It has been shown that strongest cases of co expressions are found in stable protein complexes such as ribosome or proteasome [110]. Furthermore, several studies revealed that interacting proteins are considerable more likely to be co expressed than non-interacting protein pairs [15, 209, 85]. Most importantly, gene expressions profiles of interacting proteins are co-evolving, and can be used to predict protein interactions [71].

1.4 Experimental methods for measuring protein interactions

Over the last few decades, the scientific community has created a considerable selection of different methods for measuring protein protein interactions [74]. There is a variety of approaches for analysing protein interactions. They can be genetic, biochemical, or direct physical (see table 1.1). However, they are commonly split up between high-throughput, and low-throughput methods, the first means the method can be used to screen a large quantity of interaction, and the later can not. One of the oldest methods for detecting protein protein interactions is far western blotting [229], followed by protein affinity chromatography. Another low throughput method for measuring protein protein interactions is x-ray crystallography/NMR spectroscopy [29]. The speciality of those two methods is that they provide an atomic model of the protein interface; however both of them are associated with extensive work. Furthermore, x-ray crystallography is severely limited on proteins which can be crystallized, whereas NMR spectroscopy can be applied to proteins which do not crystallise, but it comes with the cost that it doesn't create one distinct atomic model. Another technique for measuring protein protein interactions is fluorescence resonance energy transfer (FRET) which not only can measure protein interactions in real time, but also can give insight into the binding strength of the interaction [231]. This method has been recently used to generate a set of protein

interactions with known binding strength [121].

On the other hand, high-throughput methods are capable of testing vast amount of possible protein protein interactions, but they have their own distinct issues. A recent comparative assessment of interactions generated by repeated high-throughput experiments showed only small overlaps between the different experiments [108]. Furthermore, a quantitative comparison of data sets revealed that highest level of precision is achieved by combining multiple methods [219]. In this work we focused on Yeast two hybrid (Y2H), and Tandem affinity purification (TAP), for two main reasons: i) They are two most common high-throughput methods, and ii) we use their individual bias in detecting certain interaction types to further validate our predictor.

Name	high-throughput	type of interaction
Yeast two hybrid (Y2H) [63]	+	direct physical
Tandem affinity purification (TAP) [171]	+	direct physical
Protein microarrays	+	direct physical
X-ray crystallography/NMR spectroscopy [29]	-	direct physical, and structure
Far western blotting [229]	-	direct physical
Protein affinity chromatography [44]	-	direct physical
Fluorescence resonance energy transfer (FRET) [231]	-	direct physical, and binding affinity
Synthetic Lethality [118]	+	genetic interaction
co-expression [110]	+	genetic interaction

Table 1.1: Overview for experimental methods for measuring protein interactions.

1.4.1 Yeast two hybrid

As mentioned above, the yeast two hybrid (Y2H) is a high-throughput screening method for protein interactions, and it greatly accelerated the speed for measuring protein interactions. The central concept behind Y2H is that many eukaryotic transcription factors consist of two distinct domais: a binding domain (BD) which mediates binding of the promoter to the DNA, and an activation domain (AD) that activates transcription [63]. It has been shown that a transcription factor can be split into two fragments, and still maintains its biological function when both AD and BD are physically (not necessarily covalently) associated with each other. Yeast two hybrid screening of two protein A, and B works by creating artificial fused protein A-BD (also known as the bait), and B-AD (also known as prey). In the next step both of the chimeric proteins are introduced to a yeast cell, and if both of the proteins interact, the transcription (AD, and

BD) factor is formed which activates the transcription of a reporter gene (see Figure 1.5). The most common used reporter and transcription factor gene combination is the GAL4 transcription factor combined with the LacZ gene, which encodes the beta-galactosidase. There are several variations of the Y2H method: one, and three hybrid systems for detecting protein interactions with DNA and RNA, respectively [235, 62]. Furthermore, there are also mammalian and prokaryotic based hybrid approches [133, 206], and Y2H methods for measuring interactions between membrane proteins [6].

There are two general approaches for a whole genome Y2H screenings: i) matrix based , ii) and library based. In the matrix based [220] method a matrix of prey clones is generated, where each clone with a certain prey is located in a well on a plate. Then each bait strain is mated with all prey strains on an array, and when two protein interact, they are detected via expresion of the reporter gene and their location on a plate. The library based [8] approach screens baits against an unknown library of preys consisting of cDNA fragments or open reading frames (ORFs). Yeast hybrids are selected based on their survivability on specific substrates, and then the interaction proteins are determined by sequencing the DNA.

However, there are several weaknesses in the Y2H method. First of all, the interaction might not happen in yeast, since a queried protein may require a species specific folding protein, which may lack in yeast. Another issue is the fact that the whole Y2H screening takes place in the yeast nucleus, thus if the proteins are not co-localised there, the interacting proteins are found to be non-interacting (false negative). Nonetheless, Y2H has been used to measure protein protein interactions in worm [135], fly [80], and human [182].

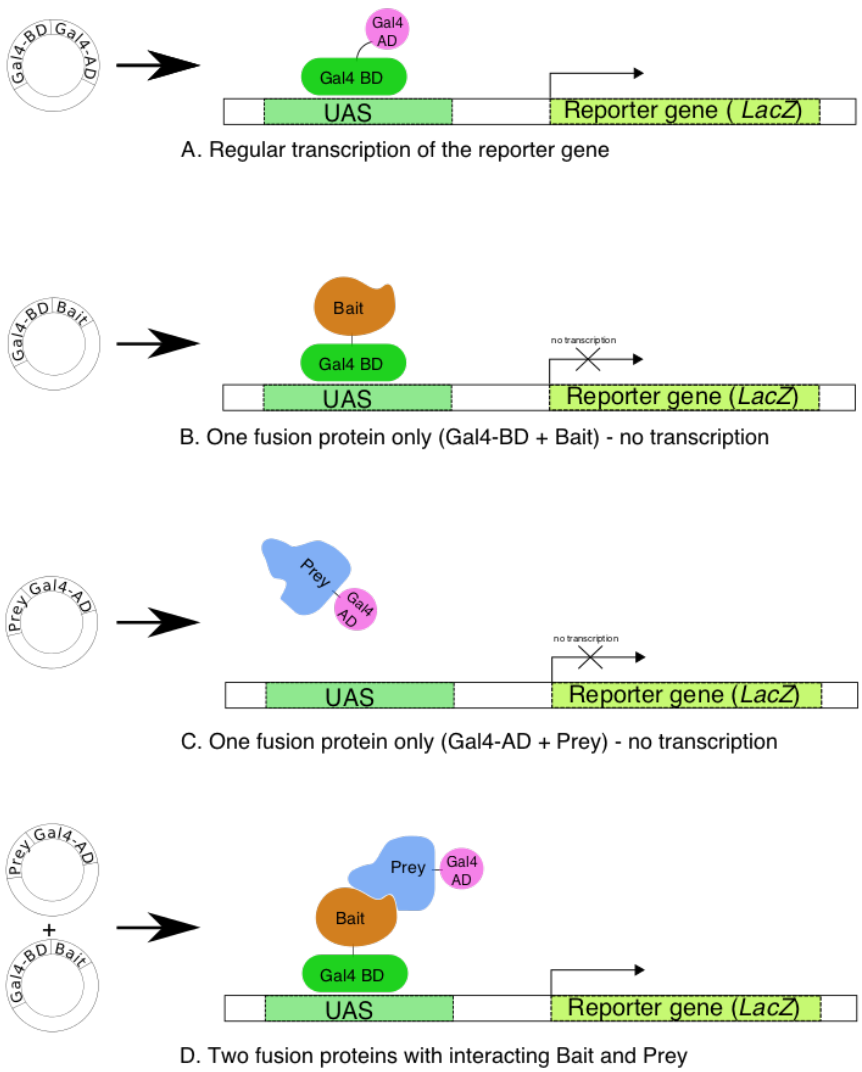


Figure 1.5: Overview of two-hybrid assay. A) In the wild type the Gal4 transcription factor enables transcription of the reporter gene. B) The prey fusion protein consisting of Gal4-BD-Bait is created and introduced to the yeast cell, however without the AD domain transcription of the reporter gene is not possible. C) Analog for the Gal4-AD-Prey fusion protein. D) When both prey and bait interact the transcription factor Gal4 is formed, and the transcription of the reporter gene starts (Figure taken from [117]).

1.4.2 Tandem affinity purification

Tandem affinity purification (TAP) is a method for rapid protein complex purification, which means the method can extract a protein with all its bound interaction partners from a cell extract [171]. In the first step of the method a fusion protein is generated consisting of the protein in question and the TAP tag. The TAP tag consists of two IgG binding domains of Staphylococcus and a calmodulin binding peptides separated by the tobacco etch virus protease cleavage site [171, 178] (see Figure 1.6). In the first step of TAP, the cell content is extracted and washed through two affinity columns. The fusion protein (TAP tag and protein of interest) bind tightly to an IgG matrix. Followed by washing out the remaining cell extract in a way that does not remove the bound interaction partes of the query protein. In the next step the fusion protein is removed from the IgG matrix via cleaving the TAP tag at the TEV cleavage site, and the protein complex subsequently gets eluted from the column. The elute is incubated with calmodulin-coated beads and calcium, which releases the target protein complex. In the last steps, the proteins of the complex are separated via gel electrophoresis, cleaved and identified via mass spectroscopy. Multiple large scale studies combine both TAP and Y2H screenings in a way to first determine direct PPIs via Y2H, and then to cluster the proteins into complexes using TAP [128].

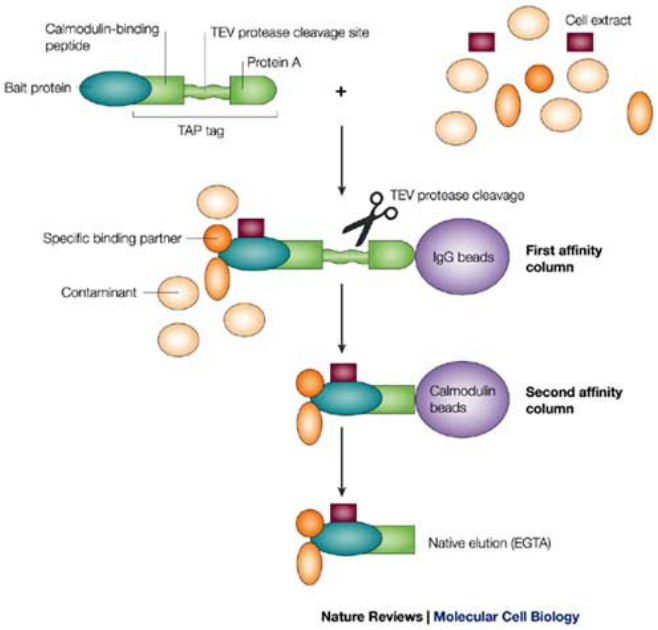


Figure 1.6: Overview of the tandem-affinity-purification method. First the artificial tap tag is created and introduced to the host organism. Cell extracts are then washed through two affinity columns. In the first column the TAP tag binds to beads coated with IgG, and then the TAP tag is broken apart by the TEV protease. In the second column the remaining part of the TAP tag binds to calmodulin beads, and is then eluted from the column via introducing calcium. (Figure taken from [104]).

1.5 Protein interaction networks

In the recent years the development and execution of high throughput methods for measuring protein interactions have caused an explosive increase in available protein interactions. Various interaction networks originated from these methods, and it became clear that a complex approach is required to understand, and quantify the topological and dynamic properties of those networks. Help was found in the mathematical construct known as graphs [86]. A graph $g = G(V, E)$ consists of a set of nodes (or vertices) V, and a set of edges E, which are tuples of nodes and they represent some kind of relation between two nodes. The edge can either be direct or undirected, where the former means that the edge has start and end point, and the later means that the edge has neither start nor end. Directed graphs are commonly used in transcription factor networks [225]; however, the in this work used PPI networks are normally consider to be undirected. There are also sever methods for measuring the network topology. The most common one is the degree of a node, which is the number of edges connecting that node to another node (aka. the neighbours of node n), and thus the degree distribution of a network is the probability distribution of these degrees over the whole network. Another property is the cluster co-

efficient, which measures how strongly a node in a graph is clustered with his neighbours. Consider three nodes A, B, and C. If all three of them are connected, it is called a triangle, and in its essence the cluster coefficient of a node n is the number of triangles that n forms divided by the total number of triangles that could pass through n. Thus, the cluster coefficient of a node n is calculated as follows:

$$C(n) = \frac{n_I}{\frac{k(k-1)}{2}} = \frac{2n_I}{k(k-1)}$$

where n_I is the number of edges connecting two neighbours of n (i.e. number of triangles), and k is the total number of neighbours n has, and $\frac{k(k-1)}{2}$ donates the upper limit of triangels n could take part in. Last but not least, a network's topology can be defined via the distribution of shortest paths between the nodes. This is captured by the betweenness (or centrality) of a node, which is the total number of shortest paths going through a node. Those different measurements (degree, betweenness, and cluster coefficient) have been used to determine various important elements of a network: such as hubs, clusters, bottlenecks. Hubs are proteins with many interaction partners and are highly connected in the network, and thus they are in general central for the connectivity of the network. Furthermore, it has been shown that there is a direct correlation between centrality of the protein in the network and how essential the protein is for the cell [112]. Clusters are highly connected subgraphs of a larger network, and typically they have more interactions within themselves and fewer with the rest of the graph [197]. Also it has been shown that clusters in biological networks form functional modules, and can also be used to assign and determine the function of unknown proteins [197]. Bottlenecks are edges which are connecting a protein in a cluster with a hub protein outside of a cluster, and thus they are key connectors of a network and a large number of shortest paths pass through them [237]. A recent study has shown that certain network motifs are enriched in protein interaction networks [150]. Network motifs are certain geometric shapes such as squares or triangels which can occur in networks. This idea was expanded by an other research groups who used network motifs to measure the local topology of a node [149], or an edge [196].

There is a huge selection of online databases for storing protein protein interaction networks. A summary of the most important databases can be found in Table 1.2. Most of the databases have different submission rules, and store a diversity of protein interaction types. For example, the DIP [186] database stores only physical interactions and allows manual submission of protein interactions. Conversely, the STRING database stores functional associations which they predict from various sources. Very popular are meta databases such as the iRefIndex [177] which collects its interactions from various other sources. More than common for databases is the option to either search and browse the database via an online interface or to download the whole database as a flat file.

Name	Number of proteins	Number of interactions	web address
DIP [186]	26743	77514	http://dip.doe-mbi.ucla.edu/
MINT [31]	35553	241458	http://mint.bio.uniroma2.it/mint/
BioGRID [33]	55296	512392	http://thebiogrid.org/
IntAct [124]	82745	291167	http://www.ebi.ac.uk/intact/
HPRD [169]	30047	41327	http://www.hprd.org/
STRING [70]	5214234	3365616679	http://string-db.org/

Table 1.2: Network sizes for various protein interaction databases.

1.6 Thesis motivation

A wide range of high-throughput experimental methods are available today for detecting protein interactions at proteome scale, but they essentially provide a binary readout — whether or not two proteins form a complex — and give no clue on how the protomers interact with each other. So far efforts to classify and predict protein interaction types have exploited structural information and are thus only applicable to the minor part of the currently known interactome for which atomic structures of protein complexes are available. In this thesis I present PiType the worlds first sequence and network based predictor for obligate/non-obligate, and SP/ME interaction types. In the following chapters I will present the results that have come from my PhD thesis.

Chapter 2

Materials and methods

In this section I will present the methods and resources used in this PhD thesis. At first I will start with the required data such as public databases, which is followed by the used data sets. In the subsequent step I will described the used features and the tools which were required to calculate them. This is followed by explaining the machine learning methods which were used to generate the features for predicting protein interaction types, as well as the used evaluation metrics. Also, I will give a detailed explanation of the methods used in the biological evaluation. Last but not least, I present the used programming languages, and the methods used to create the PiType prediction web server.

2.1 Protein sequences, structures, and annotations

Since the human genome project [214] it was clear that special places were required in oder to store, archive, and acces the flood of biological data. This lead to the development of several large scale database with the only goal to store biological information such as: Uniprot [3], and the PDB [13]. In the following sections I will explain the content of those two databases, and how they were integrated into the PiType pipeline.

2.1.1 PDB

The protein data bank (PDB) [13] was created in 1971, and is one of the first and most important databases in bioinformatics. In its essence, the PDB is a archive for macromolecular structures, who are stored as flat files with an uniform format for containing atomic co-ordinates, and partial bond connectivities information, which was derived from x-ray crystallography studies. However, macromolecular structures in the PDB are not only limited to simple one protein structures, the are multi-, singe-, homo, and hetero protein complexes, as well as, protein-protein, protein-DNA, protein-RNA

complexes, and DNA/RNA hybrids. Most importantly is the huge increase of the size of the PDB database, as a reference in 1976, 2000, and 2014 the total number of entreis in the PDB were 13, 13579, and 100843 respectively. The PDB database grows with an exponential rate, which creates several challenges on its own. On of the challenges was how to visualize protein structures, and this caused the creation of several protein structure visualisation tools such as Rasmol [187], Jmol [100], and JSmol [95]. However, the PDB had the largest impact on the field of protein structure prediction [181, 68, 21], and docking [203], as well as, making such events as the CASP [155] competition possible.

There are several ways to acces the PDB database. First of all one can reach it via their website (http://www.rcsb.org/pdb/home/home.do), or via their ftp webserver. On their website one can ether browse or search the database using keywords, or identifier. As for the ftp server it provides flat files containing the whole database as well as id mapping files which allows the mapping of PDB Ids to various other biological database, additionally the mapping can also be done using PDB SOAP web service.

2.1.2 Uniprot

The Universal Protein Resource (Uniprot) is one of the largest protein information databases, and it consist of the following subunits: the UniProt Knowledgebase (UniprotKB), the Uniprot Reference Clusters (UniRef), and the UniProt Archive (UniParc) [39]. In this work we only used the UniprotKB database, and thus I will particularly focus on explaining and describing the UniprotKB database. The UniprotKB database is the central unit of the Uniprot database, and it is a collection of sequence, and functional information on proteins, with an exhaustive and accurate selection of annotation for each protein. The UniprotKB can be split into two essential parts: one the UniProtKB/SwissProt part, which contains fully manually annotated and corrected records with additional information taken from the scientific literature and manually evaluated computational analysis, and the UniProtKB/TrEMBL part, which sonsits of automatic generated entries that await full manual annotation.

A UniprotKB entry is generated as follows. First a coding sequence is taken from one of the public nucleic acid databases such as EMBL-Bank/GenBank/DDBJ [120, 12, 153], and then it is derived and translated. All those protein sequence and related data is automatically integrated into the UniProtKB/TrEMBL database, and thus are considered to be unreviewed. In the next step Uniprot's manually annotation team critically analyse and review entries in the UniProtKB/TrEMBL, and after a verification process UniProtKB/TrEMBL entries are transfered to the UniProtKB/SwissProt part. A UniProtKB/SwissProt entry does contain all transcripts of a gene including all post translational modifications, whereas a UniProtKB/TrEMBL entry normally contains one translated

coding sequence. Consequently, a UniProtKB/SwissProt entry is generated by combining several UniProtKB/TrEMBL. Similar to the PDB database the Uniprot database can be accessed via, web interface, ftp, and soap service.

2.1.3 The gene ontology

The Gene Ontology (GO) [7] was founded as a response to the accelerating availability of genomic data [83, 214, 1, 36], and the observation that a large amount of genes have similar sequence and function [224, 122, 22, 213]. Thus, the GO adresses the need to have entirely computational system for comparing or transferring annotation among different species. Essential for this is the creation of a structured, precisely defined, common, controlled vocabulary (ontologies) for describing the roles of genes and gene products in any organism.

Ontologies are structures as a mathematic directed graph, where each node contains one GO term (e.g. DNA recombination, DNA priming, DNA helicase, etc ...), and a connection from node a to node b means that node b is a more specific version of node a (node a is parent of node b). The resulting network is a directed acyclic graph, and is best imagined as a tree where GO terms towards the root are more general and the GO terms closer to the leafs are more specific. The gene ontology can be separated into three domains: cellular component, biological process, and molecular function.

- **Cellular components:** they describe components of a cell, which are part of some larger object, such as anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
- **Biological process:** are consecutive biological events which are caused by one or more ordered assemblies of molecular functions, for example cellular physiological process or signal transduction.
- **Molecular function:** are activities, such as catalytic or binding activities, that occur at the molecular level.

The GO database can be reached via their website (http://www. geneontology.org/), and offers several ontologies for download, as well as an exhaustive collection of various tools which utilize the GO [38].

2.1.4 Integration of the Data sources

Protein sequences and associated annotation for Homo sapiens, Escherichia coli, and Saccharomyces cerevisiae were extracted from the Uniprot database [3] based on the taxon identifiers of these organisms (9606, 83333, and 559292, respectively). We only considered manually reviewed Uniprot entries to reduce the influence of wrong gene models on our

	Human	Yeast	E. coli
Reviewed proteins in Uniprot	20226	6619	4303
PDB chain IDs mapped to reviewed Uniprot entries	41605	7585	15661
Proteins with at least one mapped PDB chain ID	4519	920	1223
Proteins with at least one GO annotation	18283	5908	3744

Table 2.1: Overview of the protein information used in this work.

results. If a protein had several annotated isoforms we selected the longest one. To establish the correspondence between known three-dimensional structures and the protein sequences in our dataset we used both the Uniprot-to-PDB mapping available from the Uniprot ftp site and the PDB-to-Uniprot mapping available through the PDB [13] SOAP [163] service. The Uniprot-to-PDB mapping was reversed (i.e. converted to a list of PDB IDs corresponding to Uniprot IDs) and then merged with the PDB-to-Uniprot mapping. All PDB chain IDs that corresponded to more than one Uniprot ID were removed, but we allowed an Uniprot ID to be mapped to several different PDB chain IDs. Gene ontology [7] assignments were obtained through the QuickGO [3] proteome download page based on taxonomic identifiers. Summary statistics about protein information used in this work are shown in Table 2.1.

2.2 Dataset of obligate and non-obligate interactions

There are two well-known manually curated datasets of protein interaction types created by Zhu et al. [241] and Mintseris et al. [151, 152]. In these datasets a non-redundant set of protein complexes with known three-dimensional structure from 80 different species was classified into obligate and non-obligate (which also includes transient). However, the Mintseris dataset is not directly suitable for training our classifier as it distinguishes transient non-obligate and permanent obligate protein interactions, neglecting permanent non-obligate interactions; we do use this set for classifier evaluation (see section 3.2). The Zhu dataset was created by combining two data sources: i) a non-redundant set of protein complexes from the PDB database for which literature evidence indicates that they occur naturally and are stable as a dimer [23], and ii) a set of non-obligate interactions corresponding to protein pairs that are found in the PDB database both in the bound and unbound state [158]. In total this dataset contains 137 interactions and was used to evaluate several structure-based classifiers of protein interaction types [136, 183, 141]; however, it contains only 25 data points for human, yeast, and E. coli and is hence insufficient for our study. We therefore created a larger dataset by predicting the interaction type of E.coli, yeast and human complexes by a structure based classifier, NOXclass [241]. NOXclass employs a two-stage support vector machine (SVM) [41] algorithm to first filter out crystal artifacts and then to classify complex structures as obligate and non-obligate. The NOXclass SVM was reported to achieve the highest classification accuracy (90.9%)

Organism	Obligate interactions	Non-obligate interactions
Human	121	423
Yeast	115	55
E. coli	45	15
Total	280	493

Table 2.2: Sizes of the datasets of obligate and non-obligate interactions for different organisms.

using the following structural features: interface area, interface area ratio, area based amino acid composition, and gap volume index. For calculating the former three features NOXclass requires the NACCESS tool [103] while the latter feature is computed using SURfnet [132]. We generated a dataset of obligate and non-obligate interactions with the NOXclass predictor. A list of all structures from human, yeast or E coli with at least two chains in the biological unit was retrieved from the PDB database. Protein chains that could not be mapped to Uniprot entries were ignored. If a PDB entry contained more than two chains we considered all possible chain combinations and classified them using NOXclass. Two confidence values were obtained for each pair of protein chains - one for classifying this chain pair as a biological assembly or a crystal artifact, and another one for obligate vs non-obligate complexes. To generate our dataset we accepted only those protein chains for which NOXclass produced confidence values of at least 90% at both stages. The NOXclass predictions were subsequently merged with the manually annotated interactions from the Zhu dataset. In total we obtained 773 protein protein interactions with known or reliably predicted interaction type (Table 2.2).

2.3 Structural interaction network

The Structural Interaction Network 2.0 [125] (SIN) combines structurally resolved protein complexes into a comprehensive protein interaction network. The database was generated by first selecting experimentally determined high-confidence interactions in human and yeast from the BioGrid [33] database. Each interaction is then mapped to available PDB structures by sequence similarity. A unique feature of this resource is the classification of interactions into mutually exclusive and simultaneously possible ones. A protein interaction is said to be mutually exclusive, if two or more proteins interact with the same interface on the surface of their common partner. Otherwise, if the interactors bind at different sites of their common partner, the interaction is considered simultaneously possible. We obtained information from SIN on 3096 mutually exclusive and 816 simultaneously possible interactions in human as well as on 584 mutually exclusive and 117 simultaneously possible interactions in yeast.

2.4 Protein interaction data

In this section we describe the used protein interaction databases, which we used in this work. For the main analysis PiType we used the iRefIndex database, and for the PiType 2.0 web service we used the STRING database.

2.4.1 IRefIndex

One central problem with protein interaction data is that the required information for one species, or protein tend to be spread across multiple databases. Thus, the IRefIndex database [177] was generated with the goal to supply a unifying index that would support searching for interaction data and that would cluster redundant PPIs. This is achieved by creating a unique identifier for each protein interaction and each participant protein, using only the primary sequence of the proteins, their taxonomy identifiers. Hence, two interactions have only the same key if, and only if they contain the same set of identical protein sequences and taxonomy identifiers. Protein interactions with the same key are then considered to be redundant, and are then merged together to one entry. The IRefIndex meta-database [177] applies this methods of creating a non-redundant data set to various resources, including DIP, MINT, Intact, Biogrid, and HPRD [186, 31, 124, 33, 169].

Protein interaction data for yeast, human, and Escherichia coli were obtained by downloading the complete iRefIndex database in the PSI–MI flat file format and then we extracted the species specific interactions. Most of the IRefIndex entries uses protein identifiers from the Uniprot database, since the generated unique keys are used mostly internally (but can still be retrieved). Thus we downloaded Uniprot protein identifiers for yeast, human, and Escherichia coli from the Uniprot database, as well as mapping from secondary Uniprot IDs to primary Uniprot IDs. In turn, we used Uniprot protein information to retain only protein interactions of desired species from the IRefIndex flat file, and we mapped alle protein identifiers to their primary Uniprot ID. We considered only information on direct physical interactions measured by a variety of methods such as yeast two hybrid, tandem affinity, anti tag/bait coimmunoprecipitation, etc. An overview of the network size and the experimental data is given in Table 2.3.

2.4.2 STRING

The STRING (search tool for recurring instances of neighbouring genes) database [70] is a collection of known and predicted functional associations between proteins derived from genomic context, high-throughput experiments, co-expression, and text mining. However, The STRING database was originally intended as a public available web

	Human	Yeast	E. coli	Total
Nodes	9917	5528	2068	17513
Non-redundant interactions	41115	39045	7197	87357
Raw interactions	77742	59336	13068	150146
Yeast two hybrid (MI:0018)	13876	11055	54	24985
Anti-tag coimmunoprecipitation (MI:0007)	1311	10049	0	11360
Pull down (MI:0096)	6651	4024	78	10753
Experimental interaction detection (MI:0045)	8243	23	6	8272
Enzymatic study (MI:0415)	1002	2	4	6491
Inferred by author (MI:0363)	0	388	5898	6286
Anti-bait coimmunoprecipitation (MI:0006)	5984	22	2	6008
Tandem-affinity purification (MI:0676)	317	3631	1112	5060
Others	40358	30142	5914	76414

Table 2.3: Sizes of protein interaction networks in human, yeast, and E. coli. Non-redundant interactions: each unique combination of interactors A and B is counted as a single interaction, regardless of directionality, experimental system, and data source. Raw interactions: Each unique combination of interactors A and B, experimental system and data source is counted as a single interactiona. Each interaction detection method is annotated with its Ontology Term (e.g. MI:0018).

server for retrieving and displaying the repeatedly occurring neighbourhood of a gene [195].

Three years later in 2003 the next iteration of STRING a major shift of focus occurred which is still observable today. Instead of focusing in gene neighbourhood, STRING now predicts and evaluation functional associations [216]. Furthermore two new methods for predicting functional associations were incorporated into the STRING database: gene fusion events, and genetic profiling. Moreover, the STRING confidence score was created, which is heavily used in this, and many other works. The STRING confidence score lies between 1 and 1000 and can be created for each individual prediction method, which allows easy and direct comparison of different methods. Moreover, it can also be calculated as a combined score incorporating every prediction methods, enabling STRING to supply one easily comprehendible confidence score to each of it's associations. In its essence the STRING confidence score denotes the expected fraction of true positive functional associations. For example, a confidence score cut-off of >= 600 (i.e. take all associations in STRING with score larger or equal than 600) means that 60% of all the selected associations are expected to be true positive. The STRING confidence score was created by benchmarking each individual prediction methods' confidence score against the KEGG (kyoto encyclopedia of genes and genomes) database [119]. Basically, STRING measured for each possible prediction method score cut-off the fraction of functional associations which have the same KEGG pathway, which STRING considered to be true positive functional associations.

In 2005 the second largest change occurred in the STRING history [218]. In order to achieve larger coverage the STRING database included information transfer from one source species to another target species. This was done by using ortholog protein information from the COG (Clusters of Orthologous Groups of proteins) database [202]. Essentially, functional associations are transfered across species, if each of the two proteins of the source species' functional association have ortholog proteins in the target species.

In the following years only minor changes to the STRING database happend, such as: the creation of a easy to use web interface (including the creation of an API), the inclusion of additional prediction methods, refinement of the confidence values, as well as the orthogonal transfer method [217, 111, 200]. This leads to the final STRING version which uses seven different evidences: neighborhood, fusion, co-occurence, co-expression, experiments, database, and textmining.

We integrated the STRING database into the PiType 2.0 web service, in order to boost the number of interactions and species supported by PiType. We decided to use a larger network for classifying PPIs since scarce network data is the main limiting factor for PiType while both sequence data and functional annotation are abundantly available. We achieved this by integrating the currently most complete source of protein interactions (PPI), the STRING database (version 9.1, [70]) as the reference network to calculate network based features. We used the list of protein aliases provided by STRING (`http://stringdb.org/newstring_download/protein.aliases.v9.1.txt.gz`) in order to find correspondence between internal STRING identifiers and Uniprot entries. PiType requires Uniprot protein IDs because the FunSimMat tool used to calcu-late functional similarity also relies on Uniprot IDs. In general, Uniprot IDs consti-tute the currently most widely accepted way to refer to proteins and we thus expect that the vast majority of PiType users will utilize them. We retained only STRING IDs with one-to-one correspondence to either Uniprot-Swissprot or Uniprot-Trembl IDs. If a STRING node could be uniquely mapped to both Uniprot-Swissprot and Uniprot-Trembl, only the better annotated Uni-prot-Swissprot entry was retained. As a result we were able to map 3161134 out of the 4306670 STRING nodes (73.4%) to Uniprot (526496 ambiguous STRING nodes, with 91% of them originating from Uniprot-Trembl). Interestingly, the mapping of STRING proteins to the Uniprot database was more complete for functional associations with higher confidence scores (Figure 2.1).

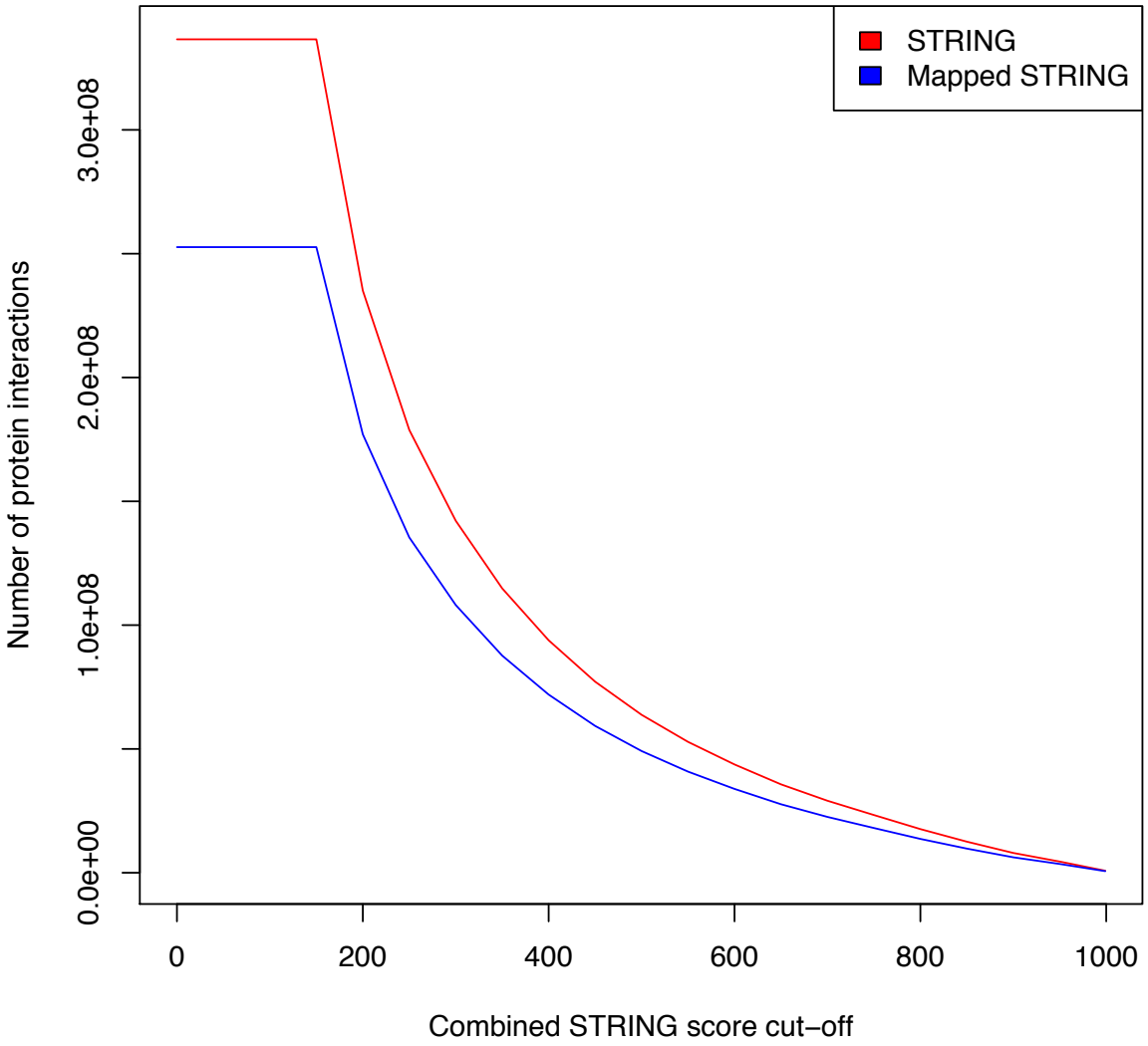


Figure 2.1: Number of interactions as a function of the combined STRING score cut-off. The red and blue lines show the number of interactions in the original STRING network and in the subset of STRING mappable to Uniprot, respectively.

2.5 Protein features used for machine learning

2.5.1 Edge graphlet degree vectors

We used edge graphlet degree vectors (EGDV) [196] as a method for measuring the local topology of an edge e in a graph g. Graphlets are small, connected, induced subgraphs of a larger network (Figure 2.2). In this work we consider graphlets of size two to five (i.e. having between two and five nodes). The local topology of an edge e can be determined by counting how often e is contained in all graphlets of size two to five in g. Moreover, one has to differentiate at which position e resides in a graphlet. For example, there are two distal edges at both ends of the graphlet G_3 and as well as one edge in the middle. To distinguish between such cases the symmetry of each edge is described by its atomorphism orbits. There is a total of 69 orbits, numbered 0 to 68. However, the orbit 0 consists of just one edge connecting two nodes. Since each edge in g touches this orbit exactly once (namely itself), it is not considered while calculating EGDV. We used a modified version of the FANMOD algorithm [226] to find all graphlets in g which contain a specific edge e (Figure 2.3) and determined the orbit of e in the graphlet using the nauty package [146]. Since values of the EGDV tend to be very large and are difficult to compare we transformed them to the natural logarithmic scale and normalized them by dividing each value by the total sum of all orbits in the EGDV (thus the sum of each orbit is 1).

We sought to identify the preferred network contexts for protein interactions of different types. To this end we investigated the enrichment of orbits in two specific local topological patterns — clusters and hubs. Edges constituting a highly connected sub-graph (cluster) would be expected to be enriched in orbits situated inside cliques, such as 2, 12, 8, 25, 52, and 68 (Figure 2.2). To be more specific orbits 2, 8, 25, 52 lie within the 3-node clique (G_2), orbit 12 within the 4-node clique (G_8), and orbit 68 within the 5-node clique ($G29$). This over-representation of certain orbits is the consequence of the large amount of combinatorial occurrences of different graphlets in tightly connected network clusters. For example, a fully connected 10-node clique, in which each node is connected to each other node, will contain the 3, 4, and 5 node sub-cliques exactly $\binom{10}{3}$, $\binom{10}{4}$, and $\binom{10}{5}$, or 120, 210, and 252, times, respectively. Thus every edge in the 10-node clique touches orbits 2, 12, and 68 exactly $\binom{10}{1}$, $\binom{10}{2}$, and $\binom{10}{3}$, or 10, 45, 120 times, respectively, and every other orbit 0 times. The lower numbers in the binomial coefficients describing orbit counts are two less in comparison to sub-cliques because for every edge the two nodes which it connects are fixed.

Clusters are also enriched in orbits (namely 8, 9, 20, 21, 24, 25, 27, 28, 51, 52, 61, 62, 63) that lie within cliques even if the associated graphlet includes further orbits that do not belong to any clique. For example, the graphlet G_6 includes three orbits — one

8 orbit and two 9 orbits — that form a 3 node clique as well as the orbit 7 which is a single attached edge to the 3 node clique. For illustration let us now consider a network consisting of 11 nodes, of which 10 nodes form a tightly connected clique, as above. In other words, an additional edge is added to the 10-node clique connecting one of the clique nodes to a node outside of the clique. In all occurrences of G6 in this network the newly added edge will correspond to the orbit 7 of G6 while the two other orbits of G_6 , 8 and 9, will lie within the 10 node clique. As a result, orbits 8 and 9 will be enriched for edges belonging to highly connected clusters.

Edges connecting hub nodes and non-hub nodes, as well as those connecting two different hub nodes, are primarily associated with orbits 1, 5, and 18. The reason for this is that hubs tend to have a high degree and a low cluster coefficient, i.e. their neighbors are sparsely connected. Edges incident to a hub are thus unlikely to form cliques.

Finally, another crucially important type of network nodes are bottlenecks, which come in two flavors: hub-bottlenecks and nonhub-bottlenecks [237]. Hub-bottlenecks are proteins characterized by high betweeness (see 2.5.3) and high degree; they are situated between protein clusters, such that a large number of the shortest paths pass through them. Nonhub-bottlenecks also display high betweeness, but their degree is low; they are the members of each respective protein cluster which interact with the hub-bottleneck node. Thus, orbits that touch a clique (such as 7, 45, 50, 57, and 58) will be enriched in interactions connecting hub-bottlenecks and nonhub-bottlenecks.

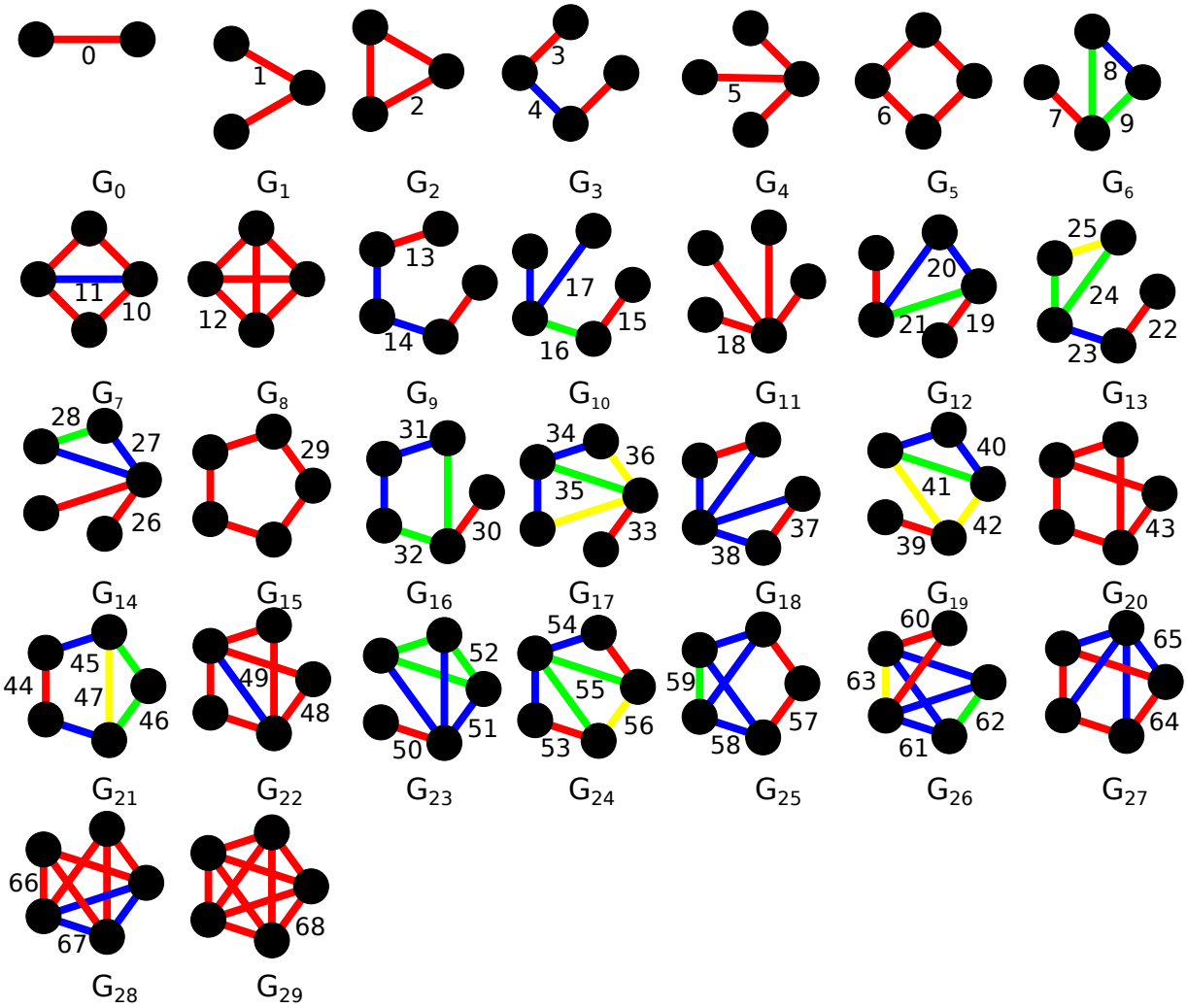


Figure 2.2: All possible graphlets of size 2 to 5 containing all 69 topologically unique edge orbits. Each unique edge orbit inside each graphlet is marked with a different color. For example, in the graphlet G13 edge orbits 22, 23, 24, and 25 are colored red, blue, green, and yellow, respectively.

```
Algorithm: EnumerateGraphlets(u, v, g, k):
Input: Edge e = (u,v), graph g = (V, E) and graphlet size k
Output: EGDV for graphlets of size k in g with edge e
    If e not in g: add(e, g), added = True
    extend = (N(u) ∪ N(v))/{u,v}
    subgraphN = N(u) ∪ N(v)
    extendSubgraph({u,v}, extend, subgraphN)
    if added: remove(e, g)
extendSubgraph(subgraph, extend, subgraphN):
    if |subgraph| == k: getOrbit(e, subgraph)
    while |extend| > 0:
        w = shift(extend)
        extend' = extend ∪ (N(w)/subgraphN)
        subgraphN' = subgraphN ∪ N(w)
        subgraph = subgraph ∪ {w}
        extendSubgraph(subgraph, extend', subgraphN')
```

Figure 2.3: Pseudo code for the EGDV calculation algorithm. N(v) denotes the neighborhood of v, i.e. all nodes that share an edge with v. A/B denotes subtraction, for example: $1, 2, 3, 5/2, 3 = 1, 5$. $A \cap B$ denotes a union of two sets, for example: $1, 2 \cup 2, 3 = 1, 2, 3$.

2.5.2 PageRank Affinity

One central problem in graph theory is to find nodes and edges which are central to the network. This problem arise both in biological [112, 90], as well as social [69], and internet [79, 130, 65] networks. Moreover, graph topology can be used to evaluate closeness of two nodes even when they are not directly connected with each other, which can provide additional insight into the network. The main idea behind PageRank affinity [215], is to use personalised PageRank [97] as measure for closeness between nodes.

PageRank Affinity is calculated as follows:
Considering a set of edges E and a set of vertices V, we define the degree of a node $u \in V$ as the number of adjacent edges, and denote it by $d(u)$. Thus we define the adjacency matrix of a graph $G = (V, E)$ as:

$$A_G(u, v) = \begin{cases} 1 & \text{if } (u, v) \in E \\ 0 & \text{otherwise} \end{cases}$$

One common method for defining graph structures are random walks. They are defined as random process where at each iteration an edge is traversed with a certain probability. Thus the transition probability matrix is the normalized adjacency matrix where each row sums to one:

$$W_G = D_G^{-1} A_G$$

and D_G donates the degree matrix:

$$D_G(u,v) = \begin{cases} d(u) & \text{if } u = v \\ 0 & \text{otherwise} \end{cases}$$

Then the PageRank rank vector $pr_\alpha(s)$, which is a steady state probability distribution of a random walk with restart probability α can be calculated via solving the linear system:

$$pr_\alpha(s) = \alpha s + (1-\alpha) pr_\alpha(s)(s)(w)$$

However Affinity PageRank always uses a starting vector that has all of its probability in one vertex:

$$e_u(i) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{otherwise} \end{cases}$$

and $pr_\alpha(e_u)$ is the steady-state probability distribution of a walk that always returns to u at restart. Thus, the personalised PageRank of a node v can be calculated as:

$$PR(v) = \sum_u pr(u \to v)$$

where $pr(u \to v)$ means the contribution that u makes to the PageRank of v.

The PageRank affinity of two nodes u, v can be defined as the minimum of the PageRank that u contributes to v, and v contributes to u:

$$PRA(u,v) = min(pr(u \to v), pr(v \to u))$$

We calculated the PageRank Affinity score using the supported java jar file, which is available for downloaded at the PageRank Affinity web server (http://gaussian.bu.edu/pnns.html). It was developed to determine whether or not two nodes share the same graph cluster. Since we expect that obligate interactions will tend to share a cluster the PageRank affinity score may be instrumental in separating non-obligate interactions from the obligate ones.

2.5.3 Betweenness

Betweenness, or centrality is a measure of a nodes's centrality in a network [72], and is defined as the number of shortest paths passing through a node:

$$g(v) = \sum_{s \neq v \neq t} \sigma_s t(v)$$

where $\sigma_s t(v)$ denotes the total number of shortest paths from node s to node t that go through the node v. In it's original definition betweenness was a node based measurement; however, the definition for betweenness of an edges in a graph is analogous (i.e. edge betweenness is the number of shortest path through an edge). Thus, an edge with a high betweenness is central for controlling the information flow and connection of the network, since removing the edge will cause a huge increase in path distance between two nodes in the network [116]. It has been shown shown that edges with a high centrality are connecting ("in between") two highly connected compounds of the network, and thus removing these edges could partition a network [81]. Furthermore, it was demonstrated that clusters in biological networks, which have been found by an edge betweenness clustering method share similar functions [56]. More recently it has been found that bottlenecks which are essential in signal transduction networks also tend to have a higher betweenness [237]. Since non-obligate interactions are typically involved in signal transduction pathways they would be expected to reside on shortest paths more frequently than obligate interactions. We used the igraph R package [43] (http://cran.r-project.org/web/packages/igraph/index.html) to calculate edge betweenness.

2.5.4 Degree

Node degree is a very common feature used in network analysis and it denotes the number of connected nodes (neighbours) of an node. The degree of an edge e is the number of edges that share at least one node with e. In other words the degree of an edge between node v_1 and node v_2 is the number of edges that have at least v_1 or v_2 as a node.

2.5.5 Eukaryotic linear motifs (ELM)

Short linear motifs (ELM) are regulatory protein modules, which are in general between three and eleven amino acids long, which generally determine the affinity and specificity of an interaction interface [156, 75]. Due to the compactness of the binding interface, more than often ELM mediate weak, transient, dynamic, and reversible protein interactions (i.e. transient interactions) [46, 131, 96]. Furthermore, it has been shown that ELMS mediate a diverse range of process such as: cell cycle progression, tagging proteins for proteasomal degradation, modulating the efficiency of translation, targeting proteins to

specific sub-cellular localisation and stabilizing scaffolding complexes [156]. Also, ELM are used by pathogens to mimic host motifs in order to regulate and control host pathways [45].

In 2003 a resource was established with the task to collect, annotate, classify, and detect short linear motifs (ELM) [173]. In its core the ELM database [50, 51] consist of two modules: i) a rational database which stores manually curated data for ELMS, and ii) a prediction method which can find all ELM occurrences in a user given protein sequence. Entries in the ELM database are grouped into ELM classes and ELM instances. An ELM class describes the specificity of an peptide binding domain (domain family), and it contains the syntax of the regularly expression for the ELM. In general an ELM class contains several ELM instances. An ELM instance contains experimentally confirmed match of an ELM classes' regular expression in a protein sequence.

Furthermore, It has been suggested that interactions in eukaryotic organisms that are mediated by short linear sequence motifs tend to be non-/obligate [166]. To determine the number of ELMs for each protein we downloaded all EKM classes from the ELM database (http://elm.eu.org/infos/news.html) and searched in each protein sequence for all occurrences of each ELM class. Hence each interaction was characterized by two integer values giving the numbers of ELMs found in both interaction partners.

2.5.6 Disordered binding regions

The central paradigm of protein structure was that a protein requires a stable, and well-defined tertiary structure. This had to be reconsidered with the discovery of so called intrinsically unstructured/disordered proteins [227, 58, 55, 207]. These proteins do not have one stable protein structure, but rather exists as a highly flexible ensemble of conformations. Disordered protein play central roles in various essential biological processes [230], such as flexible linkers between already folded domains in multidomain proteins [30]. More than often, they function via binding to other biological macro molecules (e.g. DNA, RNA, or protein) in-which they change from a disordered state to a more ordered state with stable secondary and tertiary structural elements [57]. Most importantly, disordered proteins are frequently associated with signal transduction, complex signaling and regulatory processes [212, 77]. In this work we decided to use the ANCHOR prediction server since it predicts disordered binding regions [147]. The general outline of the ANCHOR pile line can be summarised into thee steps:

1. Using the UIPred algorithm ANCHOR identifies amino acid that belong to a long disordered regions, and filters out globular domains.
2. Pairwise interaction energy in each long disordered region is calculated to make sure

that the region is not able to form enough favorable contacts with its own local sequential neighbors to fold.

3. Calculate possible binding strength with any globular domain to ensure that the region can bind with another protein.

We predicted disordered binding regions for each of the interacting protein, and considered as features the total number of disordered binding regions, the fraction of disordered amino acids, as well as the length of the longest disordered binding region in both interacting proteins. Hence, for each pair we obtained two values for the number of disordered binding regions and the fractions of disordered amino acids and one value for the length of the longest disordered binding region.

2.5.7 Functional similarity

Functional similarity between two proteins was calculated based on their associated Gene Ontology (GO) annotation [7] using the method of Wang et al. [223] as implemented in the GOSemSim package [236] (http://www.bioconductor.org/packages/2.4/bioc/html/GOSemSim.html). This method describes the similarity between two GO terms based on their location in the GO graph. To calculate the functional similarity between a protein A having G_O terms GO_1, ..., GO_i and a protein B with GO_1, ..., GO_j all i GO terms of A are compared with all j GO terms of B, yielding a matrix m with i rows j columns corresponding to GO terms of A and B, respectively. Functional similarity between A and B is then the mean over the maxima of each row and column of m:

$$funsim(A)(B) = \frac{\sum_{k=1}^{i} max(m_{i,1,...,j}) + \sum_{l=1}^{j} max(m_{1,...,i,j})}{i+j}$$

2.6 Machine learning methods

In this section we describe the applied machine learning methods. We used the Weka package [91] (http://www.cs.waikato.ac.nz/ml/weka, v. 3.6.6) and its java API for feature selection and classification.

2.6.1 Random forest

The random forest is an ensemble of several tree predictors [27]. It is calculated as follows. Given a training set of $X = X_1, ..., X_n$ and a set of class labels $C = C_1, ..., C_n$ (here ether 0, or 1), do the following steps:

1. Sample, with replacement, n training examples from X, C; call these X_b, C_b. Due to its nature of drawing with replacement, X_b, C_b contains on average only 66% of all training data.
2. train decision tree on X_b, C_b. It is import to note that for random forest each split rule is calculated using a random subset of features.
3. evaluate the created decision tree on the 33% of data point, which were not drawn in the first step, and use the evaluation result to assign the decision tree a weight α
4. repeat step 1 til 3 , B times, which is in geral between 100 and several thousand iterations, or how often user specifies

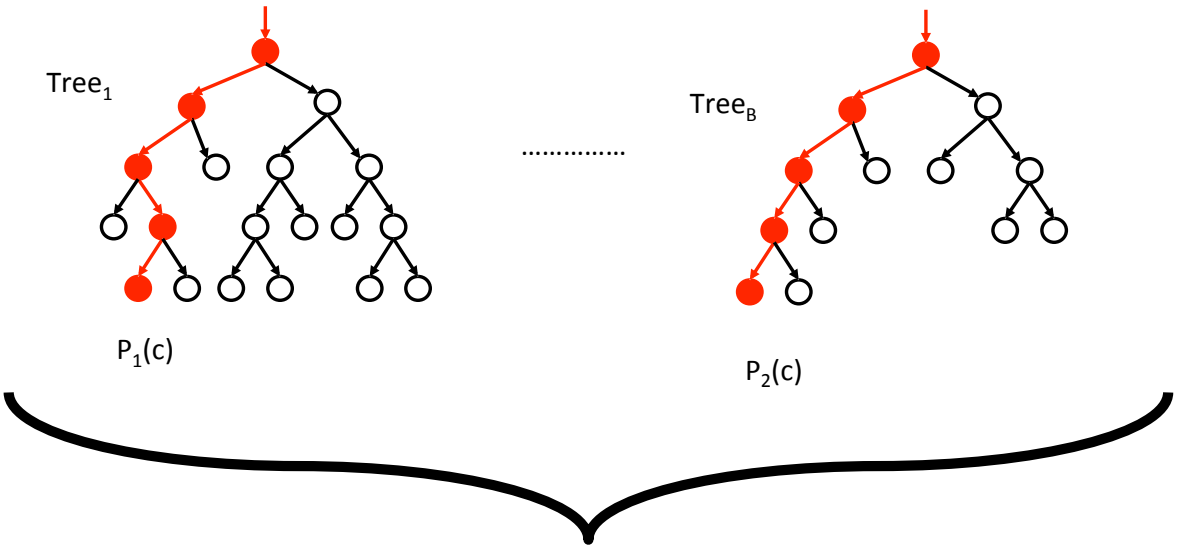
A new data point is then classified via a weighted majority vote of all created decision trees (Figure 2.4):

$$P(C) = \sum_{t=1}^{B} \alpha_t P_t(C)$$

In its essence random forest are bootstrap aggregated (bagging) [26] decision trees, with one twist and that is the random selection of features for training the decision tree. This is done to reduce correlation between the generated decision trees. We decided to use random forest, because Block et al. reported that from all tested classifiers the decision tree method achieved the best performance in distinguishing between permanent and transient interactions based on known three-dimensional structures of protein complexes [17]. Moreover, the random forest algorithm has a better accuracy and is more robust than the decision tree approach. We used the random forest classification algorithm with an ensemble of 10 decision trees. These trees are used to create a confidence value for each predicted class c, which lies between 0 and 1. This value describes the weighted fraction of decision trees that voted for class c.

2.6.2 Class balancing

One common problem in classification is that one class is over represented in the training data, here we have more ME, and non-obligate than SP, and obligate protein interaction, respectively. This problem is commonly know as class imbalance, and it became one of the central problems in the data mining community [233], since it is a common issue in several real-world classification tasks such as: fault diagnosis [234], medical diagnosis [145], or face recognition [138]. Thus several methods have been developed to solve the class imbalance problem, and depending how they resolve the issue they can be classified into three groups: A) they create, or change the algorithm in a way such that it take the class balance into account [175, 238, 228], B) they add an preprocessing step, which the data is rebalanced in oder to decrease the effect of the imbalance [9, 34], C) Combined



$$P(C) = \sum_{t=1}^{n} \alpha_t P_t(c)$$

Figure 2.4: Example of a trained random forest. The data has been used to boost several random forest and to estimate their weights α_i. In order to predict some new data point each of the decision trees are evaluated and the final prediction is an weighted ensamble of all decision trees.

methods which uses both approaches [35, 73]. A recent exhaustive comparison showed that the best performance was reached with simple approaches, which combine random undersampling techniques with bagging or boosting ensembles [76].

Thus, we chose the RUSBoost algorithm [192] for alleviating class imbalance, since it combines boosting with random under-sampling, and it has been reported that it outperforms many other approaches [192]. Boosting combines several weak learners into one strong learner. The weak learners are those trained on a subset of the training set, which is generated by random sampling with replacement. During the sampling process only two thirds of the data are used to train the weak learner while the remaining third can be used to evaluate and appropriately weight the classifier. These steps are repeated until a number of weak learners are created whose weighted combination constitutes the strong learner. The RUSBoost algorithm applies a modified resampling step. Instead of randomly subsampling the complete dataset it removes members from the majority class until a given ratio between the minority and the majority class is reached. The disadvantage of this approach is that the weak learners are always trained on the same members of the minority class. We therefore altered the RUSBoost algorithm in such a way that the members of the majority class are first removed, followed by one round of boosting on the remaining data. This process of resampling and boosting is repeated multiple times.

2.6.3 Feature selection

There are three main reasons for applying feature selection methods [185]: (a) to improve the quality of the modle, for example: classifier performance in the case of supervised classification, (b) to reduce computational time by removing unnecessary features, and (c) to gain a deeper insight into feature importance. However, since it has been reported that random forest classifier is robust to noisy data, we didn't expect any performance increase for PiType. Thus, our primary intrest in feature selection was to gain deeper insight into each feature's influence.
Essentially, feature selection methods can be divided into: wrapper, filters, and embedded methods 2.4. Filters act as a preprocessing step which removes features independently from the classifier. Wrapper, use some method to evaluate the subset feature space (e.g. classifier performance, or correlation based evaluation), and some other method to navigate the feature space (e.g. random selection, or genetic algorithm). Embedded feature selection is integrated into the classifier method, for example the random forest classifier have their own way of handling noisy data. An overview of the methods we used in this thesis is shown in Table 2.4. In the following paragraphs we will explain each of the methods, and then summarise how we used them to evaluate our features.

Information Gain:
is commonly used in deriving the split rules for decision trees [190]. In its essence the information gain IG is the mutual information $I(X;A)$ of X and A, which describes the reduction in the entropy by splitting the training cases X based on feature A:

$$IG(X,A) = H(X) - H(X|A)$$

where $H(X)$ is the entropy:

$$H(X) = \sum_{i}^{classes} P(X_i) log P(X_i)$$

, and $H(X|A)$ is the weighted entropy after the split:

$$H(X|A) = \sum_{i}^{splits} \frac{\{X \in i\}}{|X|} H(\{X \in i\})$$

In other words the information gain is the the entropy of the current state, subtracted by the weighted entropy of the two splits. Here we combine IG with ranked feature selection, which means that we always selected the top 20 features according to IG.
Correlation-based:
Correlation based feature selection is based on the principle that a good set features should be highly correlated with the class labels, yet uncorrelated to each other [92]. The merit of a subset of features S with k features is calculated as follows:

$$Merit_S = \frac{r_{cf}}{\sqrt{k + k(k-1)r_{ff}}}$$

where r_{cf} donates the average class-feature correlation, and r_{ff} is the average of all feature-to-feature correlations. Thus the optimal set of features is the one that maximises $Merit_S$. Here in this work we used a greedy forward/backward selection for finding the best feature set. Forward/backward feature selection starts with the empty feature set and in each iteration a feature is either removed or added, depending which causes the largest increase in $Merit_S$.
Genetic algorithm:
determines the best subset of features for classification via simulating biological evolution [232]. The algorithm starts by creating a population of 50 or more members, where each member is a randomly selected set of features. Then for a fixed number of generations (normally 20) the following steps are repeated:

1. **selection:** perform on each member a 10-fold cross-validation, and only retain 60% of the best members

2. **mating:** randomly select a pair of parent feature sets from the population, and randomly combine them to create a new feature set (this is repeated until the original population size is restored)

3. **mutation:** mutate the new generated feature set from the previous step

At the end the features in the best feature set (according to 10-fold cross-validation) are selected and the other features will be removed.
Random forest:
The random forest classifier we used in this work applies internal estimates to measure variable importance. Such classifier is therefore intrinsically robust even for very noisy data, so that no appreciable improvement in evaluation metrics through feature selection would be expected [27]. The implicit feature selection process is hardly traceable for the user, especially because random forest is an ensemble of multiple decision trees and one feature can be part of multiple decision rules.
Summary:
We therefore conducted a separate investigation of feature importance for each classification problem using three different methods:

1. Information gain. Features were ranked based on their information gain and the twenty best features were selected.

2. Genetic algorithm. Feature subsets were searched and created with the genetic algorithm. Each feature subset was evaluated according to the performance of the random forest classifier on the subset.

3. Correlation-based feature selection. Here the results of two rounds of greedy feature selection with correlation based evaluation - a) forward selection (start with no features and add features until convergence, i.e. when no improvement is achieved after five iterations), and b) backward selection (start with all features and remove features until convergence, as above) - are combined, retaining only those features selected in both rounds.

2.6.4 Naïve classifier

In order to have a better understanding of how well our method performs we would ideally need to compare it to other available predictors. However, since no other sequence or network-based predictors currently exist we compare our classifier to a naïve version of itself. This naïve version is a baseline RUSBoosted random forest classifier trained with only one feature. This feature is the number of proteins in an interacting pair annotated with the GO term "protein complex" or any children thereof and can thus take one of the three values 0, 1, or 2. We expect this feature to distinguish between those interacting protein pairs that are part of a protein complex and are thus presumably involved in an obligate interaction as opposed to those interactors that are not known to be part of a complex.

Name	type	advantage	disadvantage
Information gain [16]	filter	fast	ignores feature dependencies
Correlation-based [92]	filter	models feature dependencies	slower than information gain
Forward/backward selection [89]	wrapper	simple	risk of overfitting
Genetic algorithm [101]	wrapper	randomised search less prone to overfitting	computationally expensive
Random forest [27]	embedded	directly integrated in the classifier method	no insight into feature importance

Table 2.4: Feature selection method overview. In this work we use three feature selection methods: genetic algorithm, forward/backward selection with correlation-based, and information gain with ranked search (i.e. 20 best features according to information gain). In this table we list each an advantage and disadvantage of each method.

2.6.5 Performance measures

We applied nested-fold cross-validation [144] to evaluate the degree of overfitting of our predictor. Data were divided into three equal sized disjoint sets and all possible permutations of these datasets (1,2,3, 1,3,2, etc) were generated. The first set was then used to train the feature selection process, the second set to train the predictor, and the third set to evaluate the predictor. In addition to the nested-fold cross-validation we also evaluated the predictor's overall performance by a regular 10-fold cross-validation procedure. We further validated our method on a holdout set of obligate and non-obligate interactions, which involved training the classifier on the NOXclass predicted data and evaluation on the combined Zhu and Mintseris dataset. The following standard cross-validation performance measures were employed:

- Precision: the fraction of predicted interactions that are correctly classified:

$$Precision = \frac{TP}{TP + FP}$$

- Recall or True positive rate: the fraction of interactions in the evaluation set that are correctly classified:

$$Recall = True\ positive\ rate = \frac{TP}{TP + FN}$$

- False positive rate:

$$False\ positive\ rate = \frac{FP}{FP + TN}$$

- F–measure: harmonic mean of precision and recall

$$F--measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

We also plotted the receiver operating characteristic (ROC) [24] and estimated the overall performance of the classifier based on the area under the ROC curve (auROC) [94] with the help of the R package [205]. In addition we utilized Precision-Recall (PR) curves to assess the precision of our classifier for various recall values [176]. These Precision-Recall values are generated by applying various thresholds to the predictor's confidence value. Furthermore we used the R implementation of the Wilcoxon ranked sum test [102], which assesses whether one of two samples tends to have larger sample values than the other one. This test can be used to rank features with different distributions, since it is independent of the feature distribution.

2.6.6 Cross species evaluation

In order to assess the generalization power of our method across different taxonomic kingdoms we split protein interaction data into three separate datasets for human, yeast, and E. coli. We then conducted a 10-fold cross-validation of three classifiers trained on individual organism specific datasets and compared the resulting auROC values to those obtained in all possible cross species validation experiments in which a classifier trained on data from one organism is evaluated on data for another organism.

2.7 Biological validation

2.7.1 Functional enrichment analysis

The goal of the enrichment analysis is to determine whether proteins of the same class share the same molecular function, as defined by Gene Ontology [37, 11, 59, 240], more frequently than random proteins. In order to apply this approach to protein interactions the following two circumstances need to be taken into account: a) protein interactions can be both SP and obligate at the same time, since the SP/ME and obligate/non-obligate classifications are independent from each other. Thus all possible combinations of the interaction types (obligate and SP, obligate and ME, non-obligate and SP, non-obligate and ME) need to be analyzed, and b) GO annotation is only available for individual proteins and not for protein interactions. We therefore annotated protein interactions by combining the GO annotation of the two interacting proteins. For an interaction e between two proteins A and B we first retrieved all associated GO terms for protein A and protein B, and then annotated e only with those GO terms occurring in both protein annotations.

The Ontologizer tool [10] was employed to find differences in GO term enrichment between a study set and the general population of protein interactions. P-values were calculated using the Parent-Child-Union method [87]. We conducted a GO enrichment analysis for all four possible class combinations such that the population set and P-value calculation stayed the same while the study set contained interactions with the same predicted class combinations. The Ontologizer represents the enriched GO terms as a hierarchical tree.

2.7.2 Protein complex data

In this study we utilized two datasets of multi-protein complexes. One of them was the manually curated collection of 1845 human protein complexes obtained from the CORUM database [184]. Information about pairwise interactions between complex members was extracted from the iRefIndex database (see Table 2.3). We considered only those 921 CORUM complexes that formed a connected sub-graph in the iRefIndex interaction network, such that there exist a path between any two members of the complex. Furthermore we removed all protein complexes with less than 4 members, leaving us with only 244 complexes.

We also utilized the recently published study of the human protein interaction network, which revealed 13993 high-confidence interactions between 3006 proteins in HeLa S3 and HEK293 cells (further referred to as the HeLa dataset) [98]. These interactions were generated by biochemical fractionation combined with quantitative tandem affinity mass spectrometry and were further stringently filtered by an integrative computational approach, taking into account additional supporting evidence. The authors applied the ClusterOne algorithm [157] to derive 622 putative protein complexes from this network, of which 187 had already been annotated in pubic databases. Note that by design all HeLa complexes form connected sub-graphs. After the publication four proteins were removed from this dataset, reducing the total number of proteins, pairwise interactions, and clusters to 3002, 13979, and 621, respectively. We excluded from consideration 151 protein complexes with less than four members to obtain 470 HeLa protein complexes, of which 163 were previously annotated and 307 were putative, computationally derived complexes.

2.7.3 Enrichment of interaction types in protein complexes

We were interested to find out whether network clusters corresponding to protein complexes are enriched in a certain interaction type (SP/obligate, etc.) or are rather a mixture of different interaction types. Such enrichment was assessed based on the information content

of a protein complex c calculated as

$$R(c) = log_2 4 - H(c)$$

where $H(c)$ denotes the Shannon entropy

$$H(c) = -\sum_{t}^{\{type\}} P(t,c) * log_2 P(t,c)$$

and $P(t,c)$ the frequency of interaction type t in the protein complex. The information content value ranges between zero and two bits, where a value of two means that all interactions in the protein complex are of the same type and a value of zero indicates that each protein interaction type is equally represented in the protein complex.

2.8 PiType 2.0

After successfully presenting PiType to the scientific community my second part of my PhD thesis focused on creating an updated and improved version of the original PiType pipeline. We improved the method by integrating the STRING database as reference network, thus dramatically increasing the number of interactions that can be classified compared with PiType, which was based on the iRefIndex database [177]. The speed of PiType has been increased by 4-fold to cope with the much larger interaction network. Furthermore, we created a freely available web service, which is available at http://webclu.bio.wzw.tum.de/PiType/.

In order to achieve this we introduced three major changes to the PiType method due to the fact that the STRING network is several magnitudes larger than the previously used iRefIndex network. First, we modified the way the local topology of an edge is calculated, since it is the most computationally expensive part of the PiType pipeline. In order to assess the local topology of an edge we used edge graphlet degree vectors (EGDV). Graphlets are small, connected, induced subgraphs of a larger network. The local topology of an edge is measured by counting how often the edge resides at a certain location in each graphlet of the larger network. This is achieved by computing each graphlet's adjacency matrix and determining the unique location of an edge within the graphlet using the nauty package. PiType relies on a pre-computed mapping of all required adjacency matrices to edge locations, thus increasing the speed of EGDV calculation four fold. Secondly, PiType 2.0 does not consider two network features — Betweeness and affinity PageRank (i.e. a measure of closeness of two proteins) — since their calculation is no longer feasible due to the size of the STRING network. Finally, we have integrated the FunSimMat [189] Web service for calculating functional similarity between two proteins, which is one of the features required by our predictor.

2.9 Used programming languages

I had to use several different programming languages (Figure 2.5) in this thesis due to the fact that many of the used software tools/packages were designed for different programming languages. For example, iGraph [43], and GOSemSim [236] were designed for R, and Weka [91] was made for Java. Thus we decided to split PiType into three levels: I) Python scripts, which call R and calculate the required features, II) Java package which contains the actual classification of the interaction, and it serves as a interface for all other dependancies, and III) the web which was made of HTML and PHP scripts. They are interweaved as follows: the user submits a job via the web interface, the php scripts stores this information in a MySQL database, and a cron job regularly looks for new jobs to submit to PiType. In case of a new job cron executes a java jar file which calculates the necessary features (via R, and python scripts), classifies the interactions using the Weka package, and then report them back to a php script, which displays the results to the user.

Name	type	tasks
Java [5]	object oriented programming language	interface for the WEKA package [91], and as mediators between the different scripts
Python [174]	scripting language	calculate the following features: disordered, EGDV, ELM, degree. Also it was used to run the NACCESSS [103], and NOXClass [241] scripts.
R [107, 204, 78]	scripting language	calculate betweenness and GO functional similarity.
MySQL [54]	database	store network information for PiType, and submitted jobs to PiType
PHP [199]	server-side scripting language	process user input, and submit it to mysql
HTML [84]	HyperText Markup Language	used to create web pages

Table 2.5: Overview of used programming languages, which we used to create PiType, and the following PiType web service.

Chapter 3

Results and Discussion

3.1 Feature analysis

We started by searching for features (see Table 3.1 for abbreviations used) that are best discriminant for each of the two classification problems addressed in this work — obligate vs non-obligate interactions and SP vs ME interactions (Figures 3.1, 3.2) — and ranking features based on their Wilcoxon ranked-sum test P-value (for the ranked features see Tables 3.2, 3.3). For a better overview we grouped features into three distinct sets - functional similarity (BP, CC, MF, MeanSim; total of 4 features), sequence features (ELM, disorderedness, total of 8 features), and network features (degree, Affinity PageRank, betweenes, EGDV, total of 71 features).

3.1.1 Sequence based features

In this work we evaluated two sequence features — number of ELMs and number of predicted disordered regions. On average non-obligate interactions tend to have almost three times as many disordered regions than obligate interactions (rank 4 in Table 3.2, Figure 3.1a) and the proteins that participate in non-obligate interactions have a considerably higher fraction of disordered amino acids (rank 5 in Table 3.2, Figure 3.1b). Furthermore, the longest binding regions associated with non-obligate interactions tend to be twice as long as those in obligate interactions (rank 6 in Table 3.2, Figure 3.1c). These results are in line with recent reports, which state that proteins involved into non-obligate interactions tend to be more disordered than those associated with obligate interactions [59]. We also found that non-obligate interactions tend to have more ELM regions (rank 8 in Table 3.2, Figure 3.1d), which agrees with the notion that ELM primarily mediate weak transient interactions occurring in signaling [60].

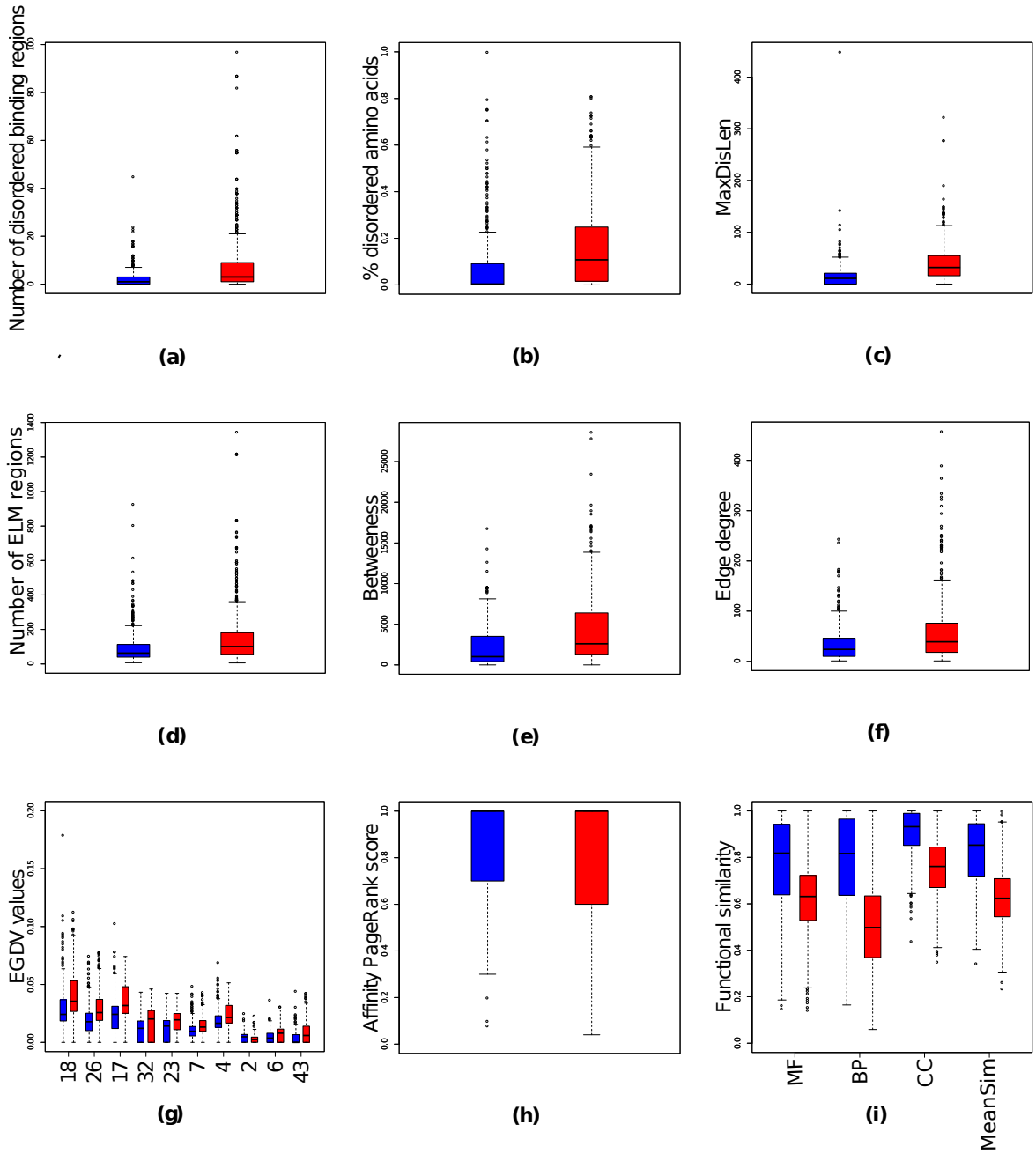


Figure 3.1: Boxplot distributions of features in obligate (red) and non-obligate (blue) interactions. For the number of disordered binding regions, the fraction of disordered amino acids, and the number of found ELM both values for protein A and B are combined into one distribution. For EGDV (g) only top 10 features with the lowest P value are plotted.

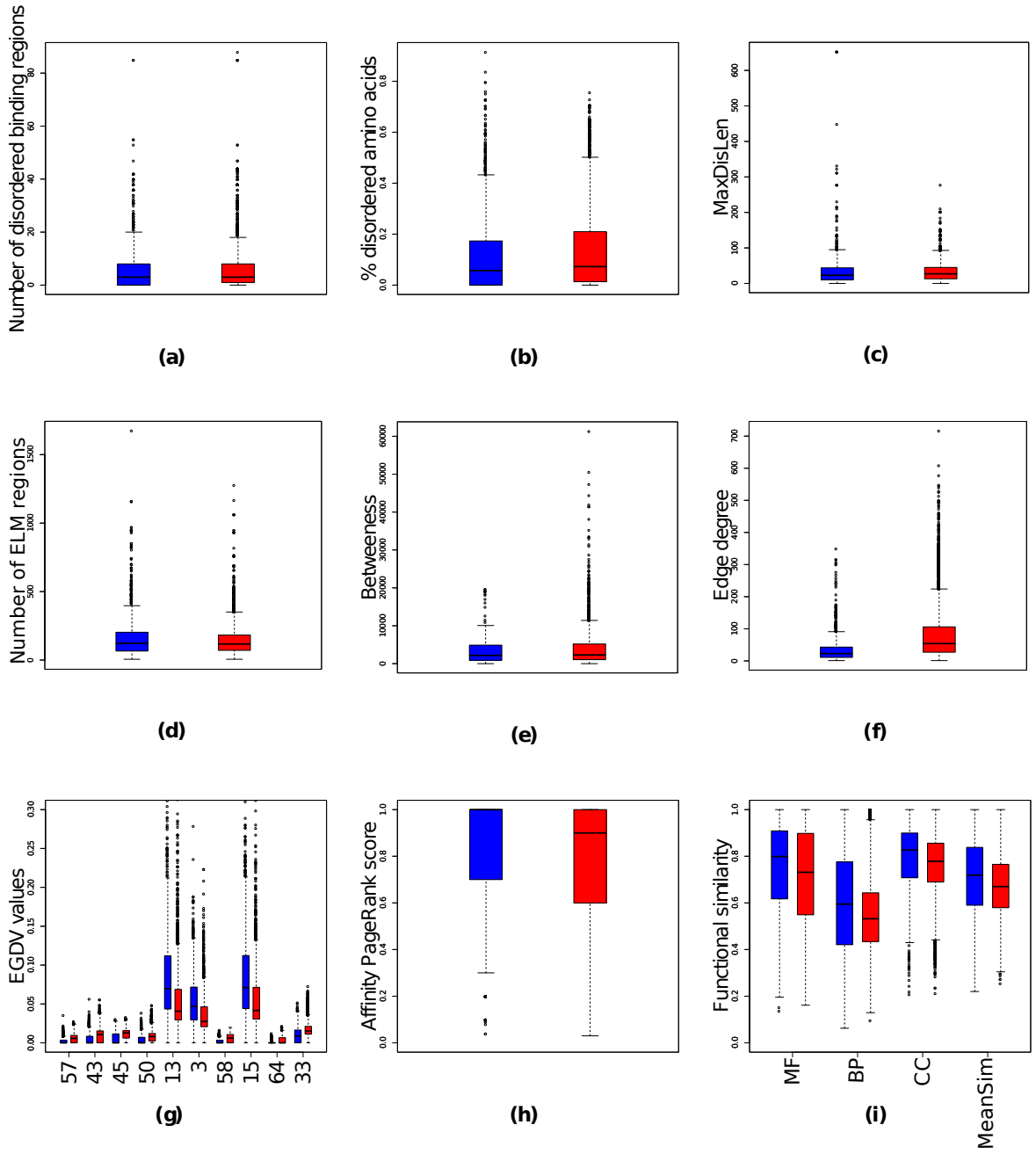


Figure 3.2: Boxplot distributions of features in simultaneously possible (red) and mutually exclusive (blue) interactions. For the number of disordered binding regions, the fraction of disordered amino acids, and the number of found ELM both values for protein A and B are combined into one distribution. For EGDV (g) only top 10 features with the lowest P value are plotted.

Name	Abbreviation
Sequence based features	
Number of found short linear eukaryotic motifs in protein A	elmA
Number of found short linear eukaryotic motifs in protein B	elmB
Number of disordered binding regions in protein A	DisRegionsA
Number of disordered binding regions in protein B	DisRegionsB
Fraction of disordered Amino Acids in protein A	FracDisASA
Fraction of disordered Amino Acids in protein B	FracDisASB
Length of the longest disordered binding regions in both proteins	MaxDisLen
Network based features	
Degree	Degree
Betweeness of the interactions	Betweeness
Affinity page rank score for the interactions	APR
EGDV values for orbit $n = 1, 2, 3, ..., 69$	1, 2, 3, ..., 69
Functional similarity based features	
Functional similarity based on cellular component GO terms	CC
Functional similarity based on biological process GO terms	BP
Functional similarity based on molecular function GO terms	MF
Mean of CC, BP, and MF values.	MeanSim

Table 3.1: Features used for machine learning

Proteins involved in SP interactions tend to be more disordered than those in mutually exclusive interactions (Table 3.3, Figure 3.2a-c), presumably because simultaneously possible interactors undergo stronger conformational changes upon binding their partners than mutually exclusive interactors [61]. At the same time we do not find any significant difference in the distribution of ELMs in SP and ME interactions (P-value 1, rank 73 in Table 3.2, Figure 3.2d).

3.1.2 Network based features

Network based features do not play a significant role in distinguishing between obligate and non-obligate interactions. Overall, they performed poorly (Figure 3.1e-f), with only betweeness showing a high rank in Table 3.2. However we do find that orbits 2, 25, 8, 52, and 68 (rank 17, 30, 31, 32, and 33 in Table 3.3) located inside cliques are enriched in obligate interactions while orbits 18, 26, 17, 32, and 23 describing hub-like proteins (Figure 2.2, section 2.5.1) are enriched in non-obligate interactions (rank in 10-14 Table 3.3, Figure 3.1g). In particular, the orbit number 68, the five clique, occurs three times more often in obligate interactions than in non-obligate interactions (P-value 0.00001, rank 33 in Table 3.3), yet the signal is too weak to distinguish those classes efficiently. This observation is compatible with the fact that obligate interactions are permanent

Rank	Feature name	Mean obligate	Mean non-obligate	P-value	Rank	Feature name	Mean obligate	Mean non-obligate	P-value
1	MeanSim	0.816	0.63	7.2e-48	21	33	0.0095	0.0148	2e-08
2	CC	0.9	0.748	1.7e-47	22	45	0.0049	0.00841	6.5e-08
3	BP	0.768	0.517	2.7e-41	23	50	0.0040	0.00737	8.6e-08
4	DisRegions	2.17	7.27	9.3e-39	24	57	0.0016	0.00395	2.7e-07
5	FracDisAS	0.0737	0.163	2.4e-35	25	30	0.0224	0.0281	5.3e-07
6	MaxDisLen	17.5	42.2	1.9e-31	26	Degree	35.4	59	7.7e-07
7	MF	0.779	0.626	4.3e-24	27	31	0.015	0.0216	2.1e-06
8	ELM	96.2	151	1.3e-18	28	39	0.0155	0.0193	3.9e-05
9	Betweeness	2463	4315	4.8e-16	29	12	0.0027	0.00132	0.0001
10	18	0.0293	0.04	6.5e-14	30	25	0.0225	0.0148	0.0001
11	26	0.019	0.0274	7.8e-13	31	8	0.0122	0.0082	0.0001
12	17	0.0239	0.0326	7.4e-12	32	52	0.0079	0.00432	0.0001
13	32	0.0109	0.0177	4.2e-11	33	68	0.0017	0.00065	0.0001
14	23	0.0122	0.0181	4.3e-11	34	51	0.0065	0.00398	0.0007
15	7	0.0105	0.0141	2.1e-09	35	41	0.0082	0.00484	0.0008
16	4	0.0173	0.0222	6.3e-09	36	20	0.0298	0.0353	0.001
17	2	0.0043	0.00258	6.9e-09	37	24	0.0216	0.0159	0.0017
18	6	0.0045	0.00752	8e-09	38	44	0.0058	0.00858	0.0022
19	43	0.0037	0.00804	1e-08	39	11	0.0031	0.00161	0.0026
20	5	0.0267	0.0323	1.2e-08	40	49	0.0027	0.00114	0.0028
41	9	0.0112	0.00883	0.0053	61	40	0.0124	0.0107	1
42	37	0.0104	0.00701	0.0098	62	53	0.00665	0.00561	1
43	67	0.00194	0.00086	0.013	63	34	0.0123	0.0107	1
44	35	0.00705	0.00456	0.013	64	21	0.012	0.0105	1
45	28	0.0192	0.0135	0.023	65	38	0.0074	0.00666	1
46	63	0.00238	0.00118	0.028	66	27	0.0149	0.0133	1
47	19	0.0196	0.015	0.04	67	54	0.00514	0.00469	1
48	APR	0.864	0.778	0.04	68	58	0.0032	0.00369	1
49	66	0.00278	0.00149	0.04	69	22	0.0468	0.0441	1
50	29	0.0126	0.0152	0.047	70	42	0.00853	0.00793	1
51	14	0.0389	0.0446	0.049	71	48	0.00473	0.00525	1
52	62	0.00329	0.00178	0.072	72	56	0.00281	0.00283	1
53	61	0.00406	0.00273	0.076	73	64	0.00164	0.00155	1
54	59	0.00317	0.0016	0.16	74	3	0.0541	0.0504	1
55	55	0.00447	0.00309	0.18	75	13	0.087	0.0791	1
56	65	0.00218	0.00133	0.26	76	47	0.00414	0.00414	1
57	1	0.0313	0.0231	1	77	36	0.0103	0.0098	1
58	46	0.00907	0.00727	1	78	60	0.00305	0.00296	1
59	10	0.00553	0.00472	1	79	15	0.082	0.0785	1
60	16	0.0564	0.0591	1					

Table 3.2: Ranked features for the obligate and non-obligate classes based on their Wilcoxon ranked sum test P-values. The numbers in the name column refer to EGDV values for orbits (see Table 3.1). For the number of disordered binding regions, fraction of disordered amino acids, and ELM both values for protein A and B are combined into one distribution which has two values for each interaction. The Wilcoxon P-value is then calculated for this distribution.

Rank	Feature name	Mean obligate	Mean non-obligate	P-value	Rank	Feature name	Mean obligate	Mean non-obligate	P-value
1	Degree	36.6	86.9	2.8e-106	21	42	0.00631	0.0108	8.2e-45
2	57	0.00207	0.00584	2.5e-95	22	16	0.0657	0.0502	1.2e-44
3	43	0.00486	0.0104	9.8e-86	23	47	0.00333	0.00622	2.7e-44
4	45	0.006	0.0113	5.2e-82	24	10	0.00372	0.00619	1.5e-43
5	50	0.00407	0.00853	3.6e-81	25	36	0.00772	0.0125	1.3e-4
1 6	13	0.0895	0.0552	9.5e-79	26	65	0.00099	0.00263	7.6e-38
7	3	0.0556	0.0363	1.2e-76	27	32	0.0145	0.0206	1.4e-31
8	58	0.00211	0.00563	1.4e-76	28	66	0.00116	0.00263	5.9e-29
9	15	0.0871	0.0558	3.8e-76	29	34	0.00912	0.0136	1.7e-26
10	64	0.00076	0.00315	3.9e-74	30	38	0.00576	0.00866	3.4e-26
11	33	0.00999	0.0166	2.1e-66	31	55	0.00267	0.00481	2.3e-23
12	48	0.00309	0.00716	3.7e-66	32	23	0.0152	0.0197	1.1e-22
13	54	0.0033	0.00707	3.7e-65	33	61	0.00211	0.00397	6e-22
14	22	0.0491	0.0344	2.8e-63	34	40	0.00921	0.0133	9e-21
15	60	0.00176	0.00444	9.9e-61	35	59	0.00186	0.003	9.8e-21
16	56	0.00187	0.0048	7.1e-58	36	7	0.0117	0.0144	5.6e-20
17	6	0.00608	0.00917	9e-53	37	62	0.00172	0.0029	2.1e-19
18	44	0.00651	0.0107	2.4e-50	38	67	0.00076	0.00156	1.1e-17
19	1	0.0265	0.0199	3.4e-48	39	14	0.0474	0.039	2.4e-17
20	53	0.0044	0.00793	2.3e-46	40	46	0.00712	0.00968	2.2e-16
41	63	0.00113	0.00199	6e-16	61	27	0.0129	0.016	8.4e-05
42	51	0.00355	0.00557	1.1e-13	62	MF	0.75	0.715	0.00024
43	12	0.00135	0.00187	1.2e-13	63	29	0.0147	0.0168	0.00052
44	11	0.00194	0.00259	2.9e-13	64	28	0.0162	0.0167	0.00076
45	49	0.00134	0.00213	3.1e-12	65	MaxDisLen	35.4	33.1	0.0049
46	35	0.00488	0.0071	9.7e-12	66	30	0.0287	0.0268	0.0092
47	52	0.0046	0.00609	9.7e-12	67	4	0.022	0.0215	0.068
48	CC	0.793	0.764	1.4e-11	68	DisRegions	5.55	5.28	0.07
49	31	0.02	0.0243	1.6e-11	69	8	0.0104	0.0101	0.4
50	26	0.0227	0.0273	2.4e-11	70	Betweeness	3476	3916	1
51	68	0.00055	0.001	4e-11	71	2	0.00662	0.00315	1
52	39	0.0165	0.0196	4.6e-11	72	19	0.0163	0.0182	1
53	MeanSim	0.714	0.677	7.3e-11	73	ELM	166	140	1
54	APR	0.825	0.742	1e-09	74	24	0.0178	0.0178	1
55	37	0.00783	0.00903	7.3e-09	75	20	0.0339	0.0344	1
56	41	0.00584	0.00733	1.2e-08	76	25	0.0201	0.0175	1
57	FracDisAS	0.116	0.136	3.6e-08	77	9	0.00931	0.0106	1
58	5	0.0321	0.0298	5.9e-08	78	17	0.0315	0.0325	1
59	21	0.0104	0.0137	1.9e-07	79	18	0.0372	0.038	1
60	BP	0.599	0.554	7.9e-07					

Table 3.3: Ranked features for the SP and ME classes based on their Wilcoxon ranked sum test P-values. The numbers in the name column refer to EGDV values for orbits (see Table 3.1). For the number of disordered binding regions, fraction of disordered amino acids, and ELM both values for protein A and B are combined into one distribution which has two values for each interaction. The Wilcoxon P-value is then calculated for this distribution.

and usually occur in functional modules corresponding to tightly connected clusters in interaction networks [62]. Indeed, we observed slightly larger APR values for obligate interactions (Figure 3.1h) that for non-obligate interactions.

In contrast, network topology differs greatly between SP and ME interactions, since there is a physical limit to how many interaction partners can simultaneously bind to a protein [22]. While we observed no significant difference for betweeness (P-value 1, rank 70 in Table 3.2, Figure 3.2e), degree is the best feature to separate these two classes (rank 1 in Table 3.3, Figure 3.2f). There are also differences in local topology, with the orbits 13, 3, 15, 22, and 6 enriched in SP interactions and orbits 57, 43, 45, 50, 58 being more prominent in ME interactions (Figure 3.2g). ME interactions were enriched in orbits describing bottlenecks, with 58 being the only exception (Figure 2.2, section 2.5.1), which implies that ME interactions are key connectors in the interaction network and that at least one of the two interacting proteins has a higher chance to be an essential gene [40]. In contrast, SP interactions prefer sparsely connected orbits (Figure 2.2, section 2.5.1), probably due to the physical limits of binding multiple partners simultaneously. Similar to obligate interactions, SP interactions tend to have larger APR values (Figure 3.1h).

3.1.3 Functional similarity

For the obligate/non-obligate classification the most significant P-values were reached with functional similarity features (ranks 1-3 in Table 3.2, Figure 3.1i). This is caused by the fact that all obligate interactions are permanent while the majority of non-obligate interactions are transient [4]. The only permanent non-obligate interactions are antibody-antigen and enzyme-inhibitor interactions. Further work is needed to distinguish those interactions from signaling and receptor-ligand interactions, which would open up the possibility of classifying interactions as permanent or transient and also distinguishing between strong and weak interactions.

For the SP/ME classification we only find a weak correlation with functional similarity (ranks 48, 53, 60 62 in Table 3.3, Figure 3.2i). However, all functional similarity features had a P-value of less than 0.05 (significant level). Thus, we expect it to play at least some part in the classification.

3.2 Performance evaluation

As an additional evaluation method we trained a random forest classifier (RUSBoost ratio 0.37 for obligate/non-obligate, and 0.31 for SP/ME, see 3.2.1) with either all features

or only with features from each individual group (functional similarity, network based features, sequence based features). In a 10-fold cross-validation the auROC values for obligate/non-obligate classification were 0.881, 0.810, 0.822, and 0.772 for all features, for functional similarity, sequence features, and network features, respectively.

Analogously, we performed the same analysis for simultaneously possible and mutually exclusive interactions. The random forest auROC values (RUSBoost ratio 0.31) were 0.851, 0.657, 0.806, and 0.808 for all features, for functional similarity, sequence features, and network features, respectively. Using either disordered features or ELM features separately we achieved an auROC of 0.75 and 0.66, respectively. However, when both disordered and ELM features were utilized the auROC was substantially higher — 0.806 — underlying the importance of cross-talk between these two groups of biological properties.

3.2.1 Class balancing

Both datasets used in this work for training the classifiers are characterized by strong class imbalance. Specifically, in the obligate/non-obligate dataset the fraction of obligate interactions is 0.20 while in the SP/ME dataset the fraction of SP interactions is 0.36. We plotted precision and recall values for each interaction class for various fractions of the minority class (obligate, SP) in the training set (Figure 3.3), since we were interested in how the class balance affects the classifier performance. With increasing fraction of the minority class in the training data the classifier precision for the minority class declines, but its recall for this class increases; the opposite tendency was observed for the majority class (SP, non-obligate). We therefore sought to identify the optimal values of the two parameters of the RUSBoost method - the fraction of the minority class in the training set and the number of iterations - and to evaluate their influence on the classification results.

Overall, the random forest classifier was very robust with respect to different fractions of the minority class (Figure 3.4). Only when minority class instances constituted less than 10% or more then 90% of the data a severe effect on the auROC could be observed. Thus, according to auROC analysis, any value between 0.10 and 0.90 is acceptable. As for the F-measure (Figure 3.5), the obligate/non-obligate classifier was stable for fractions of obligate interactions between 0.40 and 0.60 while for the SP/ME classification F-measure values peaked between SP fractions of 0.30 and 0.40. As for the overall performance of the classifier (Figure 3.5, green line) we observed that for obligate/non- obligate classification the F-measure peaks at a value of 0.811 for the class ratio of 0.37, and for SP/ME classification the highest F-measure value of 0.829 was reached with the ratio of 0.31. As an additional test we trained one random forest classifier with 100. trees, with all features selected and no class balancing, and obtained

F-measure values of 0.804 for obligate/non-obligate interactions and 0.811 for SP/ME interactions, respectively, evaluated in a 10-fold cross-validation experiment. This implies that utilization of these class ratios for the RUSBoost method can slightly improve the F- measure of both classifiers. Thus in this work, for all further analyses, we decided to use the ratio of 0.37 for the obligate/non-obligate classification and 0.31 for the SP/ME classification. Even though we only achieved a small improvement through class balancing, we were able to demonstrate the robustness of our classifier for a broad range of class ratios.

With regard to the number of RUSBoost iterations no significant improvement could be achieved for more than five iterations as judged by the auROC values (Figure 3.6). We therefore decided to use 10 iterations throughout this work to ensure optimal performance.

3.2.2 Predictor evaluation

In this subsection we describe performance evaluation of our random forest classifier for both obligate/non-obligate and SP/ME classification in comparison with the corresponding naïve classifiers, serving as baseline. The naïve classifier achieves an auROC value of 0.69 for the obligate/non-obligate classification using protein complex annotation information while for the SP/ME classification its performance is essentially random (auROC=0.54). Table 3.4 shows performance measures for each class in a 10-fold cross-validation. We observed higher precision, recall, and F-measure values for the majority classes (non-obligate, ME), than for the minority classes (obligate, SP). The reason for this is the fact that these evaluation metrics are susceptible to the class imbalance in the data set. This is also the reason why we obtained the lowest values for the SP classification, since only 21% of the SP/ME dataset contains SP interactions. In terms of the overall performance, our classifier achieved the auROC values of 0.881 and 0.851 for the obligate/non-obligate and the SP/ME classification, respectively.

Table 3.5 shows the results of the 10-fold and nested-fold cross-validation for obligate/non-obligate classification, both with and without feature selection. For all feature selection methods we observed a minor decrease in performance compared to no feature selection. Information gain feature selection performed better than genetic and correlation-based feature selection, despite having the largest standard deviation. The results of the 10-fold and nested-fold cross-validation and holdout set analysis are quite similar, which demonstrates classifier robustness (Figure 3.7a,b and Figure 3.8a,b). There is also a significant improvement over the naïve classifier, justifying our use of additional features compared to the naïve version. However, there are noticeable jumps on the precision-recall curve for the holdout set (Figure 3.8a,b), presumably due to the small sample size.

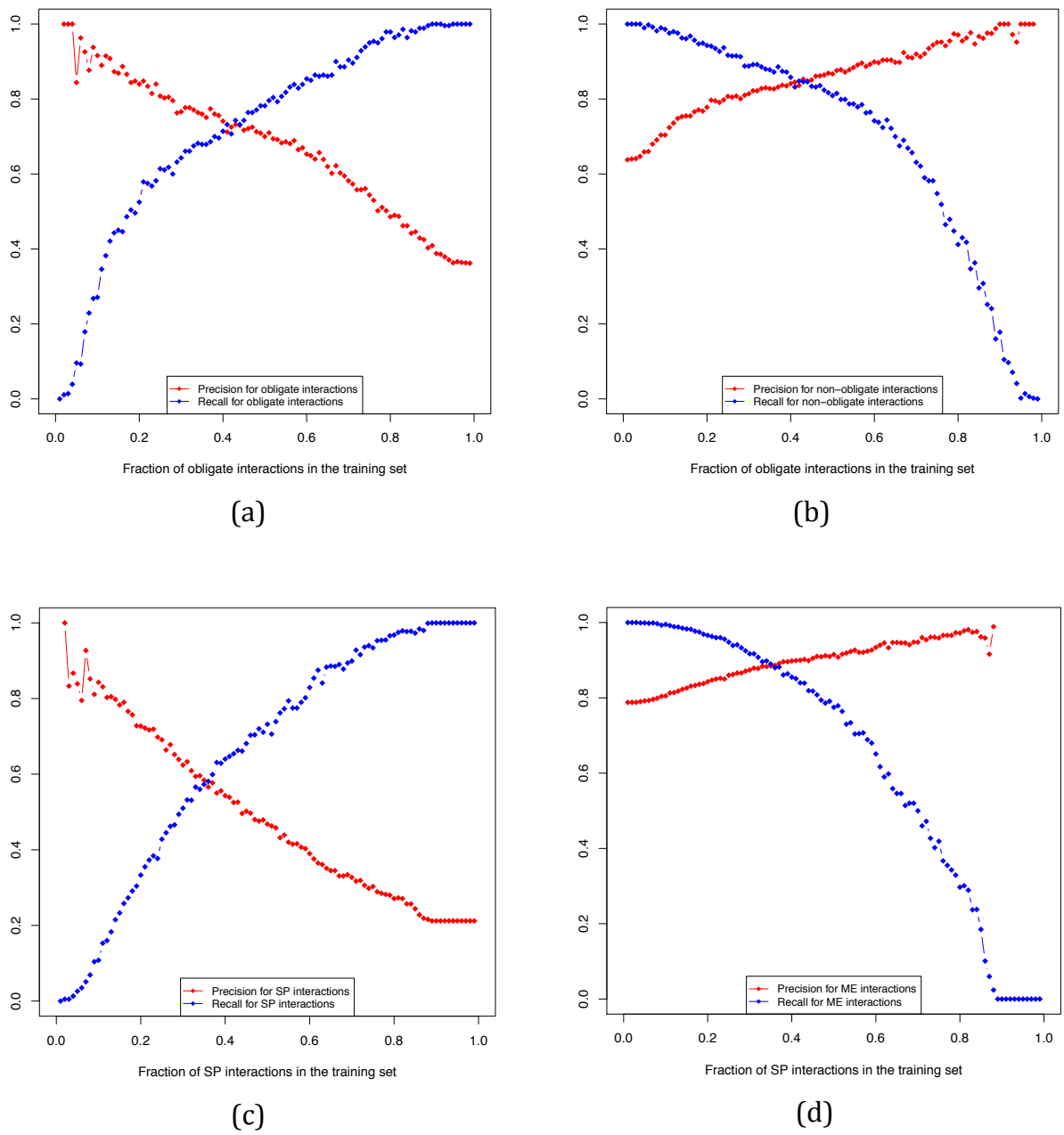


Figure 3.3: Precision and Recall values for different fractions of obligate (a), non-obligate (b), SP (c), and ME (d) interactions in the training set.

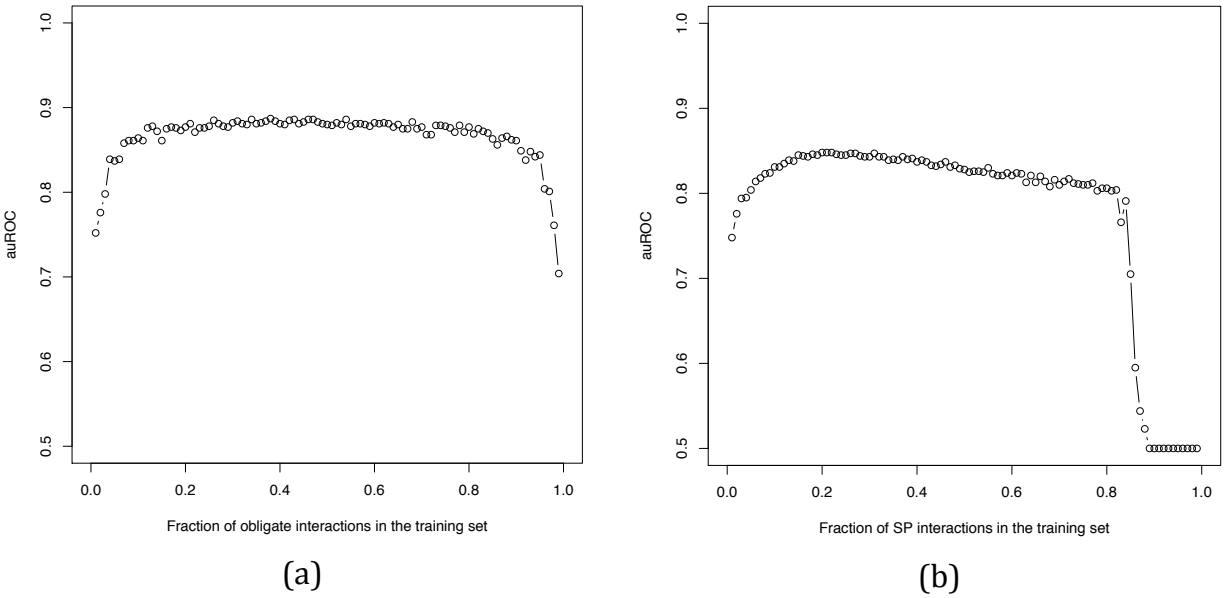


Figure 3.4: AuROC of the classifier for different fractions of obligate (a) and SP (b) interactions in the training set.

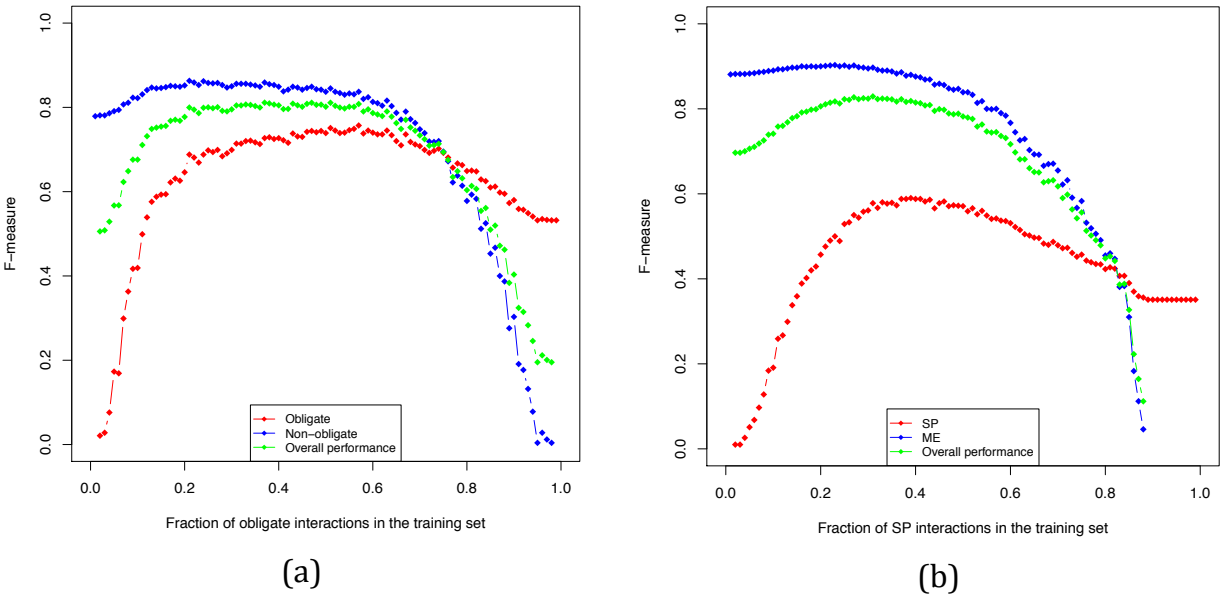


Figure 3.5: F-measure of the classifier for different fractions of obligate (a) and SP (b) interactions in the training set.

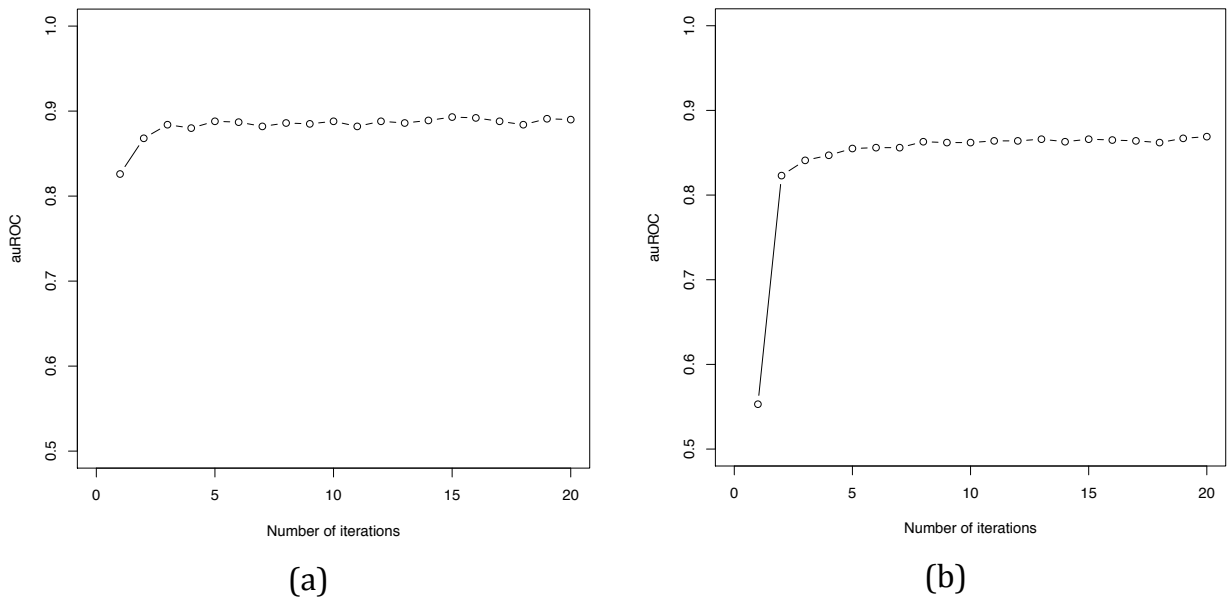


Figure 3.6: AuROC for different numbers of RUSBoost iterations.

We were primarily interested in the results of the correlation-based feature selection since it selected only 10% of all features (MF, BP, CC, meanSim, betweeness, FracDis-ASB, and MaxDisLen) at the cost of an essentially negligible decrease in performance. For obligate/non-obligate interactions, we observed a difference of 0.02 in auROC and F-measure between 10-fold cross-validation and nested-fold cross-validation. The deviation within the nested-fold cross-validation was merely around 0.01 (Table 3.5).

For SP/ME classification, correlation-based feature selection outperforms all other methods in 10-fold cross-validation, even without feature selection, while in nested- fold cross-validation it outperforms only all other feature selection methods (Table 3.6). The selected features are: 26, 32, 33, 43, 50, 57, 64, degree, elmA, elmB, BP, and MaxDisLen. The fact that "functional similarity based on biological process GO terms" was selected by the correlation based feature selection method, serves as evidence that BP has at least some importance for the classification. We observed only small deviations between 10-fold cross-validation and nested-fold cross-validation and within nested-fold cross-validation (Figure 3.7c,d and Figure 3.8c,d).

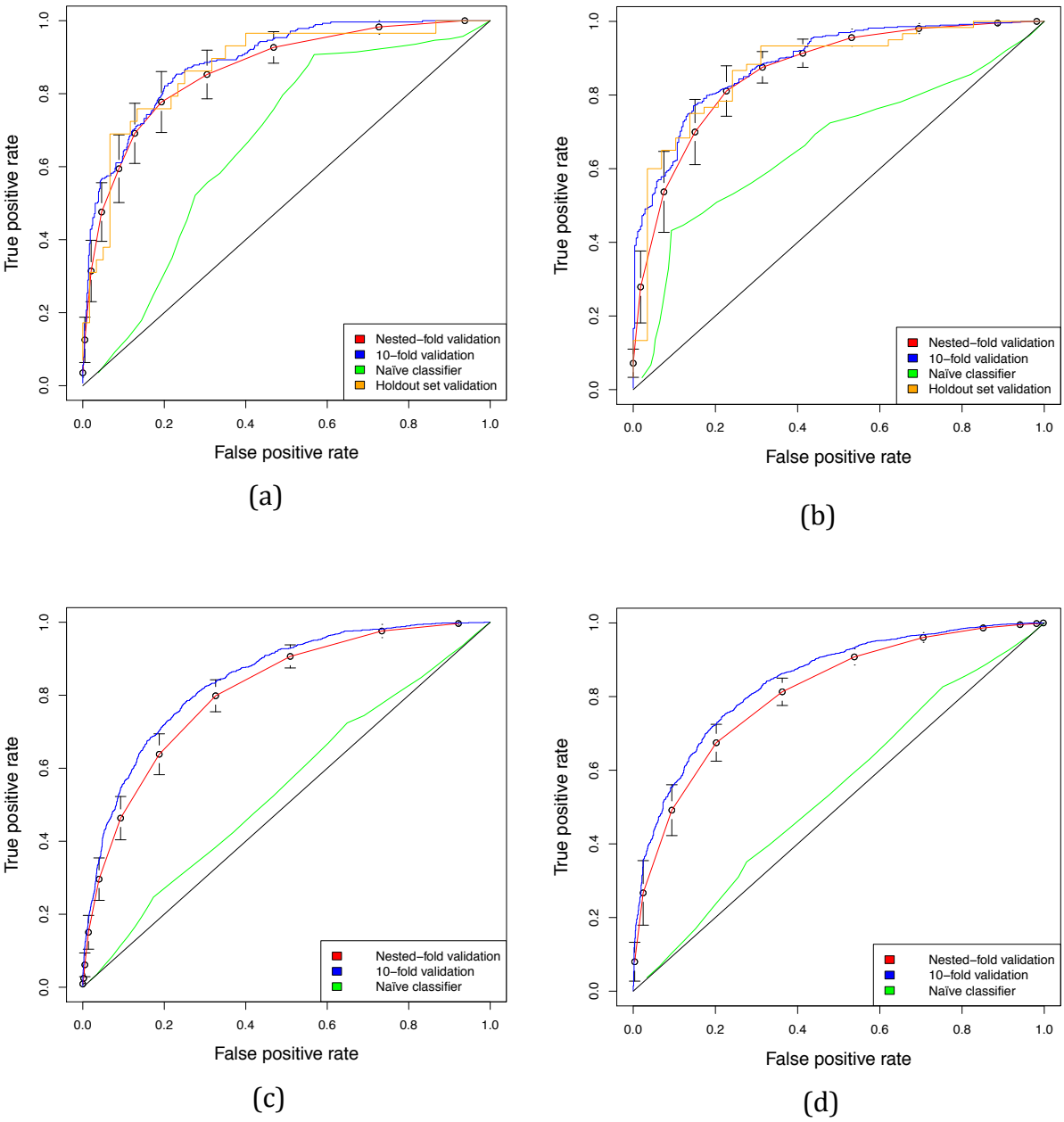


Figure 3.7: ROC curves for obligate (a), non-obligate (b), SP (c), and ME (d) classification.

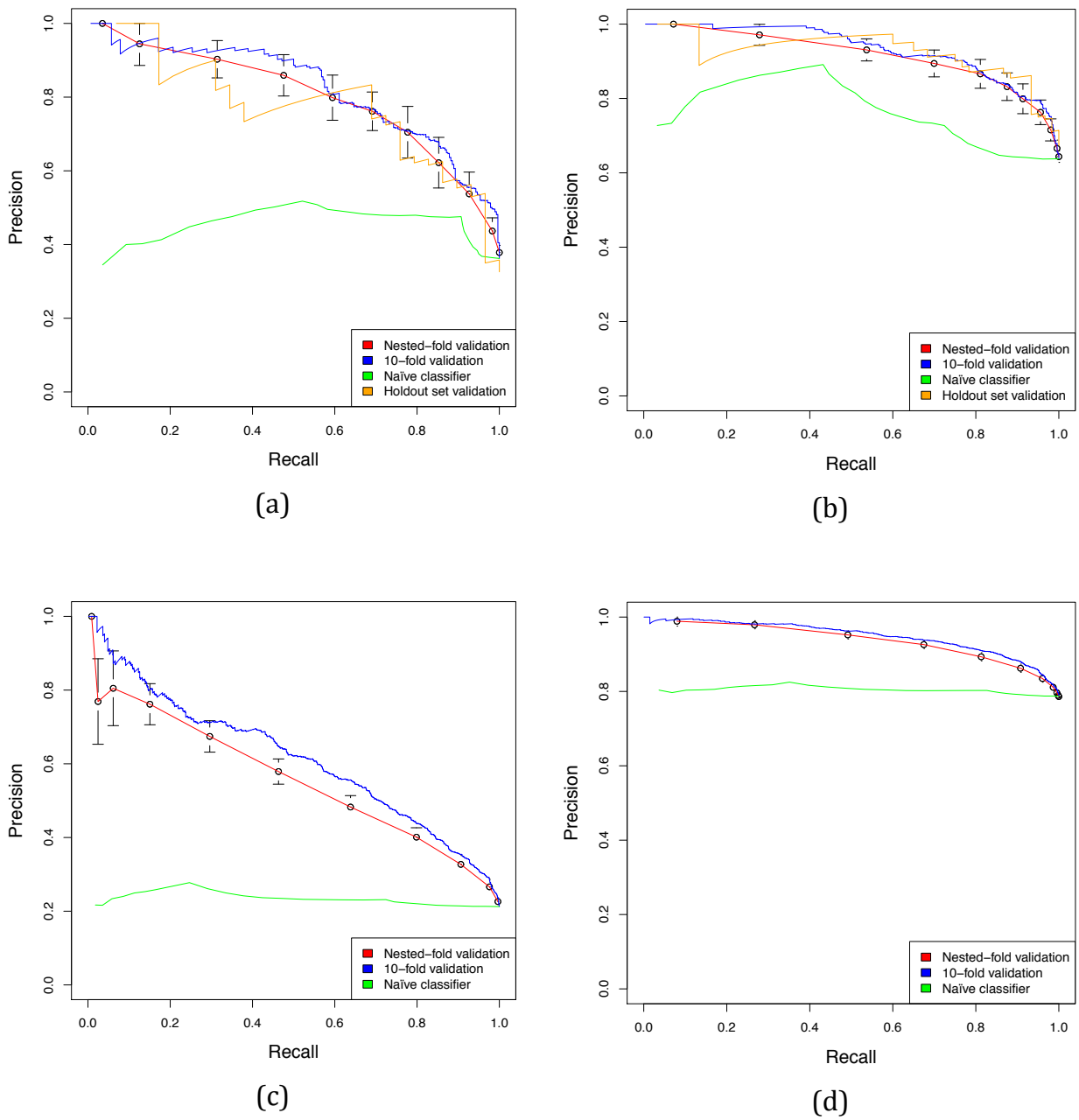


Figure 3.8: Precision/recall curves for obligate (a), non-obligate (b), SP (c), and ME (d) classification.

	Obligate	Non-obligate	SP	ME
Precision	0.75	0.83	0.61	0.87
Recall	0.68	0.87	0.53	0.91
F-measure	0.71	0.85	0.56	0.88
auROC	0.881		0.851	

Table 3.4: Evaluation metrics for obligate, non-obligate, SP, ME classification for 10-fold cross-validation. auROC describes how well the classifier can distinguish both classes, hence there is only one value for each classifier (obligate/non-obligate, SP/ME).

	10-fold cross-validation			Nested-fold cross validation		
Feature selection method	Number of features	F-measure	auROC	Number of features	F-measure	auROC
No feature selection	82	0.811	0.881	82	0.794±0.009	0.869±0.015
Information gain	20	0.803	0.877	20	0.804±0.017	0.871±0.019
Correlation	8.7	0.807	0.865	9.33	0.797±0.012	0.863±0.014
Genetic	38	0.808	0.880	35.6	0.793±0.017	0.869±0.015

Table 3.5: Evaluation of feature selection methods for the obligate/non-obligate classification. The "Number of features" column refers to the average number of selected features in each fold. For nested-fold cross fold validation the standard deviation for each value is given with the "±" symbol. Performance measures (auROC and F-measure) reflect the overall performance of the classifier.

	10-fold cross-validation			Nested-fold cross validation		
Feature selection method	Number of features	F-measure	auROC	Number of features	F-measure	auROC
No feature selection	82	0.829	0.851	82	0.802±0.008	0.808±0.003
Information gain	20	0.783	0.791	20	0.769±0.004	0.769±0.011
Correlation	14.1	0.81	0.837	13.6	0.794±0.009	0.812±0.011
Genetic	41.7	0.837	0.816	37	0.795±0.013	0.811±0.011

Table 3.6: Evaluation of feature selection methods for the SP/ME classification. The "Number of features" column refers to the average number of selected features in each fold. For nested-fold cross fold validation the standard deviation for each value is given with the "±" symbol. Performance measures (auROC and F-measure) reflect the overall performance of the classifier.

	Human	Yeast	E. coli
Human	0.842/0.832	0.750/0.824	0.808
Yeast	0.768/0.718	0.814/0.766	0.731
E. coli	0.726	0.719	0.761

Table 3.7: Cross species evaluation for SP/ME (left number) and obligate/non-obligate (right number) classification. For E. coli only one number for obligate/non-obligate classification is shown since no SP/ME data is available for this organism. Presented are auROC values of the classifiers trained with the data from the organisms shown in table rows and evaluated on species shown in table columns. Diagonal values (same species used for training and evaluation) were derived by 10-fold cross-validation. The off-diagonal elements show cross-species evaluation where the classifier was trained on the row species and evaluated on the column species.

3.2.3 Cross-species evaluation

In this section we evaluate the performance of our classifier trained on data from one organism and then tested on data from another organism. The goal of cross species evaluation is to determine whether or not we can apply our classifier to species other than human, yeast, and E. coli. Classifiers trained and evaluated on data from the same organism have larger auROC values (Table 3.7, diagonal elements) than those obtained in all possible cross-species validation experiments (Table 3.7, off-diagonal elements). The only exception is constituted by the obligate/non-obligate classifier trained on human data and evaluated on E. coli, which has a somewhat larger auROC than the classifier trained and evaluated on E. coli data. This might be caused by the fact that the E. coli obligate/non-obligate dataset (60 interactions) is considerably smaller than the human obligate/non-obligate dataset (545 interactions). In general, we see that for obligate/non-obligate interactions classifiers trained on individual organism-specific datasets perform only marginally better on these native datasets (difference in auROC values between 0.01 and 0.05) than on data from organisms they have not been trained on. Interestingly, for SP/ME interactions the difference between organism-specific and cross species evaluation is slightly higher (between 0.05 and 0.1), but is still quite acceptable. This might have been caused by the fact that the obligate/non-obligate dataset was generated by the structure-based NOXclass classifier and is thus more homogenous. In conclusion these results suggests that both the obligate/non-obligate and SP/ME classifiers can be applied to analyze interaction data from species other than the three organisms considered in this work, albeit with a slightly larger decline (up to 0.1) in auROC for the SP/ME classifier.

3.2.4 iType 2.0 web server evaluation

Upon introducing changes into the original PiType pipeline concerning the reference network for calculating EGDV and degree features we conducted extensive benchmarking,

since the STRING network includes not only physical interactions, but also functional associations. We determined the optimal cutoff value of the combined STRING score based on 10-fold cross-validation on the our SP/ME, and obligate/non-obligate training data [82] with the following STRING score cutoffs: 400, 450, 500, 550, 600, 650, 700, 750, 800, 950, and 999. We observed essentially the same precision, recall, F-measure, and auROC values for obligate/non-obligate classification (Figure 3.9) as for PiType. The ME classification was also as accurate as before, while for SP we observed a decrease in precision for higher STRING cut-offs and an increase in recall for higher STRING cut-offs. F-measure for SP classification achieves its maximum for the STRING cut-off of 950. We therefore decided to use 950 as the optimal STRING score cut-off. With this setting, the performance of PiType 2.0 is on par with the original pipeline (Table 3.8) while allowing for classification of a considerably larger amount of interactions. Furthermore, our classifier shows comparable performance both for physical protein interactions obtained from the iRefindex database and for the STRING network, which is primarily based on functional associations rather than direct physical contacts between proteins. However, we found that highly confident STRING associations are strongly enriched in experimentally confirmed physical contacts. For example, for the combined STRING score cut-offs of 400 and 950 the fraction of STRING associations with a non-zero score based on either experiments or databases is 33% and 82%, respectively (Figure 3.10). Thus, the highly confident STRING network with the cut-off of 950 mostly consists of physical protein interactions, which explains why the classifier performance is similar to the original pipeline.

The user of the Web server can select the desired features to be used for classification and supply a list of interacting protein pairs as well as a target species (Figure 3.12). The server accepts Uniprot, Ensembl [66], and RefSeq [170] protein identifiers as well as Entrez Gene's GeneID identifiers [140] and gene names. Non-Uniprot identifiers are automatically mapped to Uniprot IDs of the target species. If no organism name is supplied, the server displays a list of species arranged in descending order by the number of mapped proteins. The user is then given the opportunity to select a species before submitting the query. On average PiType 2.0 requires one minute to classify one network edge. The number of interactions a user can submit at once is currently limited to 1000 (approximately one day of runtime); larger networks can be processed upon request. When the server has finished classifying the interactions, the user is automatically redirected to the result page containing a list of unmapped proteins followed by a table displaying one interaction per line as well as confidence values for SP, ME, obligate, and non-obligate classes. Confidence values for each predicted class, produced by the random forest classifier, reflect the fraction of decision trees that voted for the respective class.

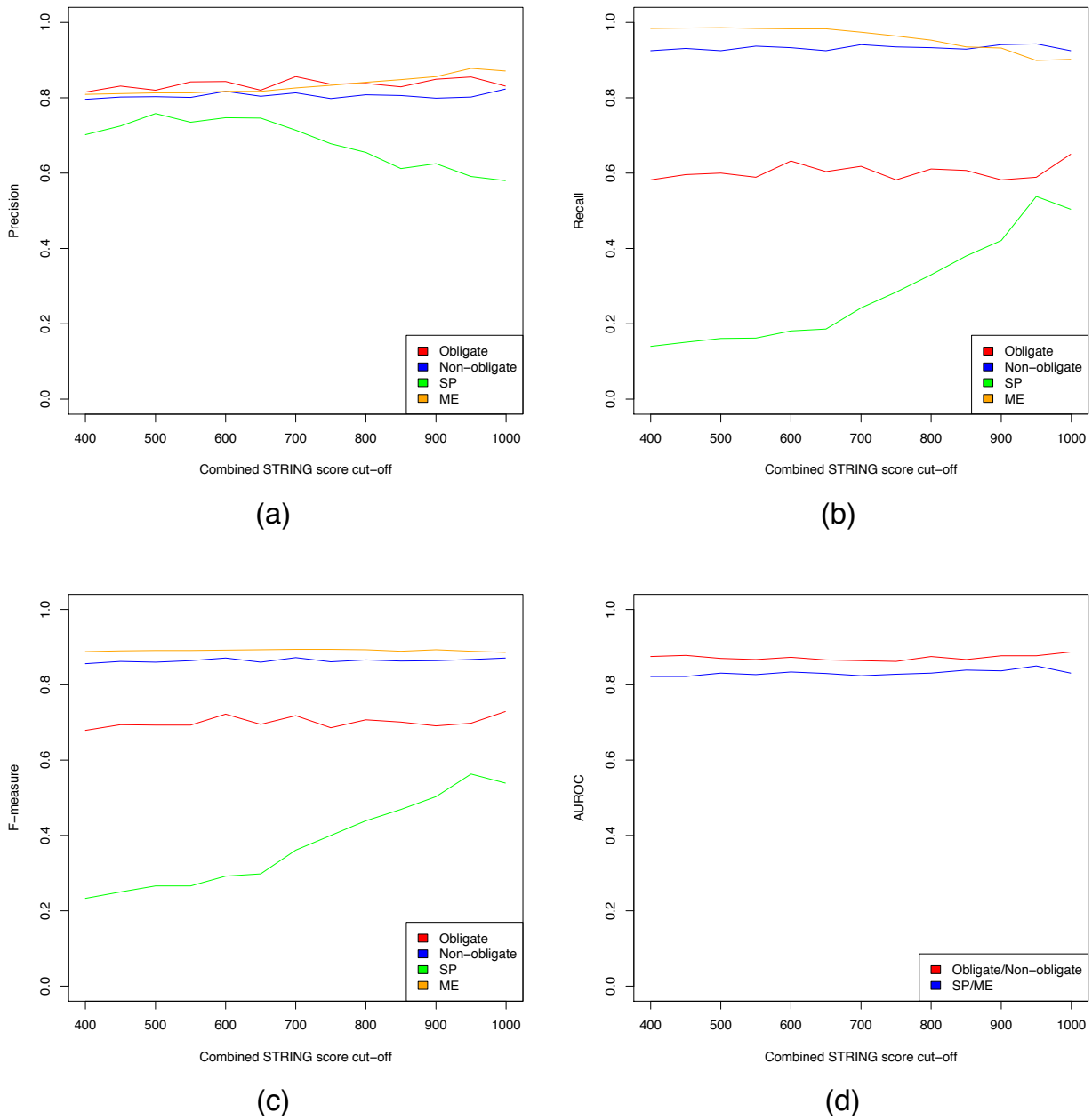


Figure 3.9: Dependence of precision (a), recall (b), F-measure (c), and auROC (d) on the cutoff value of the STRING combined score for obligate, non-obligate, SP, and ME classification.

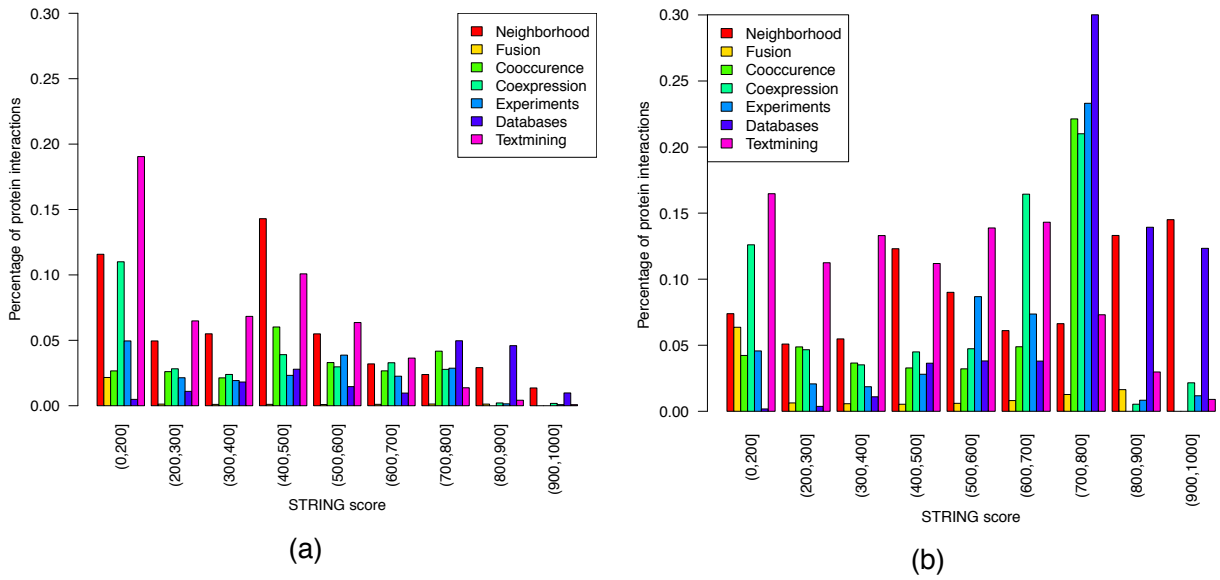


Figure 3.10: (a) Distribution of STRING scores for experimental evidence (Experiment and Databases) for the STRING interaction network with a combined STRING score cut-off of 400. The interactions are binned according to the confidence of each evidence, where "$(900, 1000]$" donates the left-closed, right-open interval between 900 and 1000. (b) The same for the STRING interaction network with a combined STRING score cut-off of 950.

	Precision	Recall	F-measure	auROC
Obligate	0.85	0.59	0.70	0.877
Non-obligate	0.80	0.94	0.86	
SP	0.59	0.53	0.56	0.850
ME	0.87	0.89	0.89	

Table 3.8: Evaluation metrics for obligate, non-obligate, SP, and ME classification with 10-fold cross-validation for PiType with STRING as reference network. AuROC describes how well the classifier can distinguish both classes, hence there is only one value for each classifier (obligate/non-obligate, SP/ME).

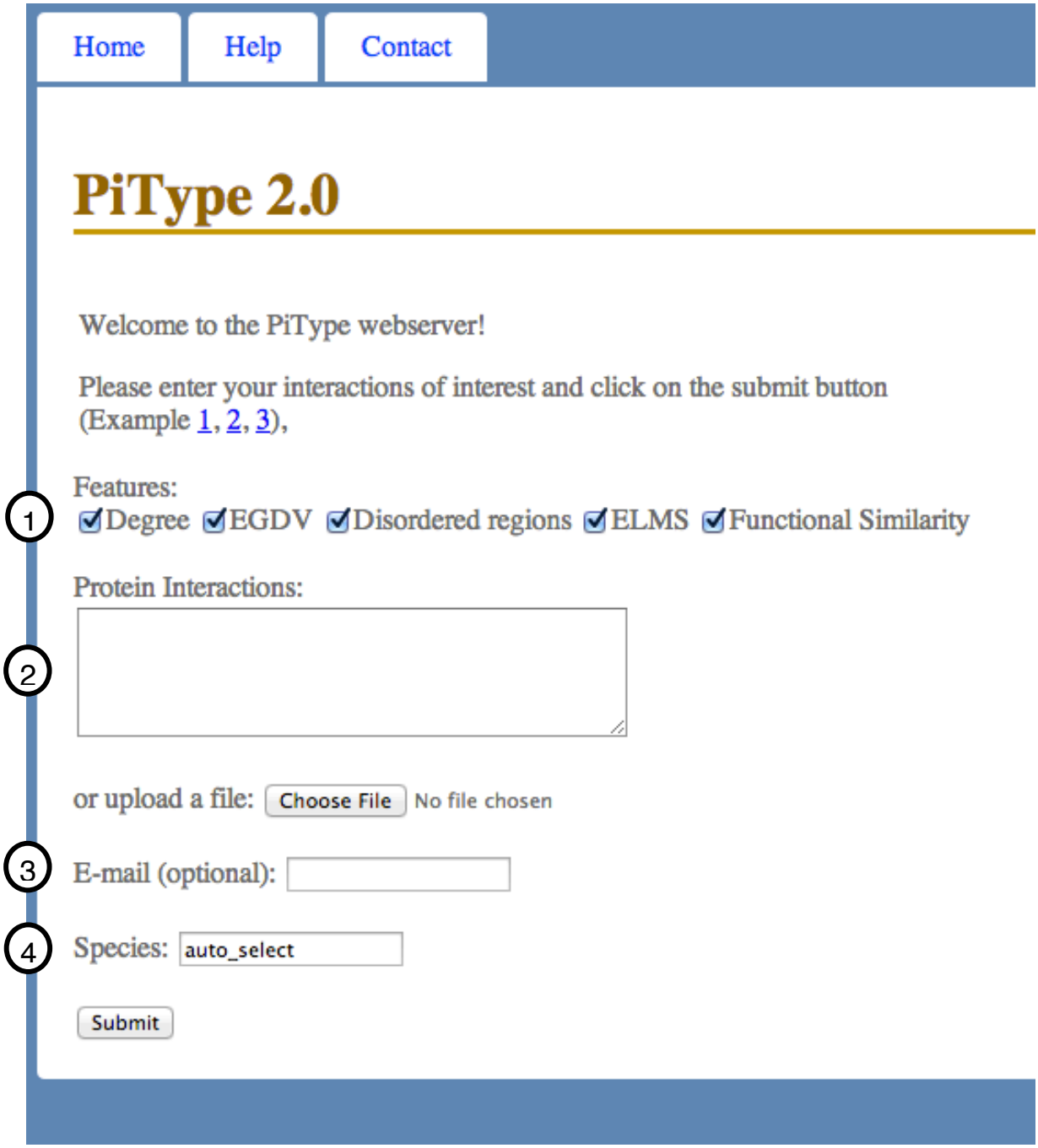
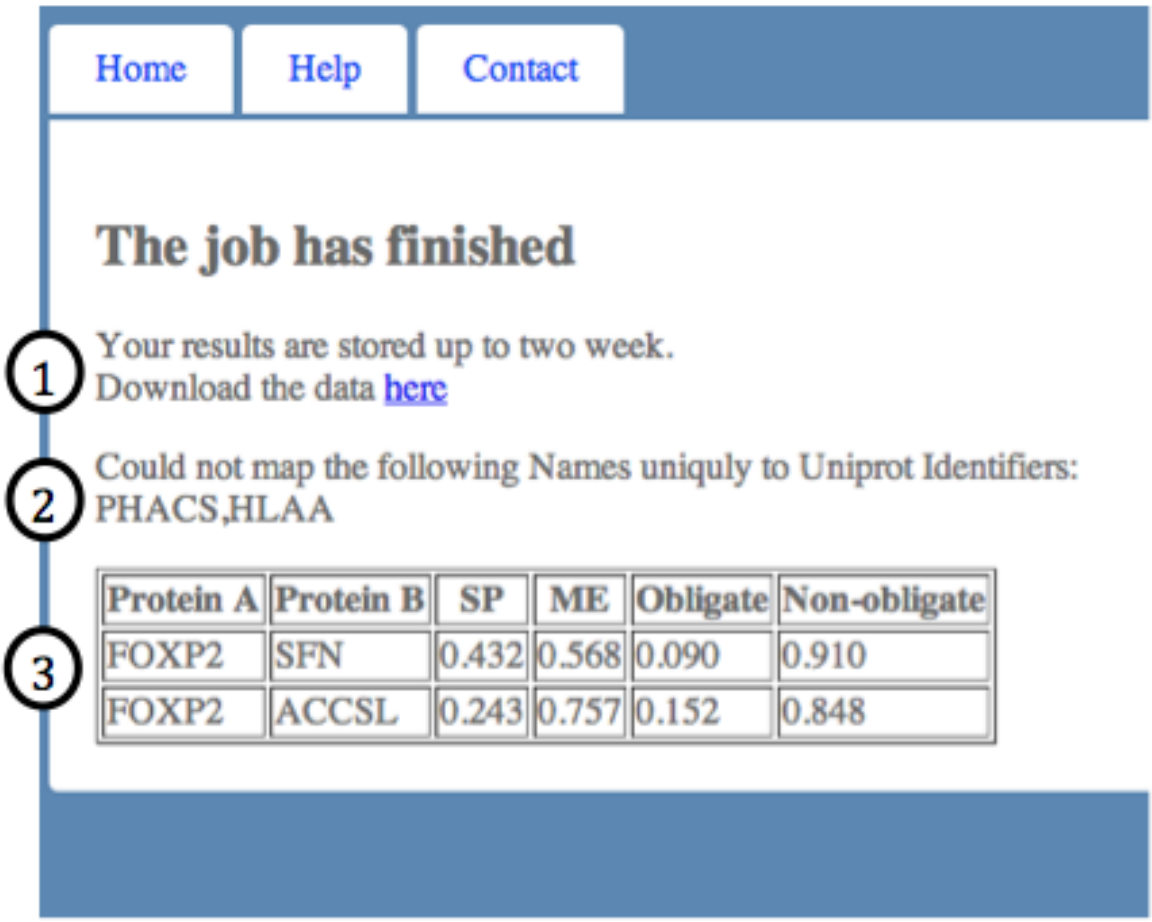


Figure 3.11: Overview of the PiType 2.0 web server interface. The input fields are as follows: 1) feature selection, 2) list of interacting proteins (either flat file or text field), 3) optional email address, 4) species (default auto_select).



Protein A	Protein B	SP	ME	Obligate	Non-obligate
FOXP2	SFN	0.432	0.568	0.090	0.910
FOXP2	ACCSL	0.243	0.757	0.152	0.848

Figure 3.12: Overview of the PiType 2.0 web server output. The fields are as follows: 1) Download server output as flat file, 2) Lists all Protein IDs which could not be mapped, since there is no mapping or they are ambiguous, 3) Table containing classified interactions and their confidence values.

3.3 Large scale classification of protein interactions

We applied our method to classify 13978 HeLa and 83788 iRefIndex protein interactions as either obligate or non-obligate as well as either SP or ME. Each interaction was also attributed to one of the four class combinations — obligate and SP, obligate and ME, non-obligate and SP, or non-obligate and ME — and assigned two confidence values - one for the SP/ME classification and one for the obligate/non-obligate classification.

We analyzed the number of classified interactions for each class and class combination for various random forest confidence values (Figure 3.13, 3.14). Note that we ignored cases where the classifier was indecisive (i.e. confidence 0.5 for both classes). The number of classified cases declines with increasing stringency of the classifier. For example, at the random forest confidence value of 0.7 two thirds of interactions get classified and around 10% are left at the 0.9 threshold.

What is the optimal threshold for the random forest confidence values? As seen in Figure 3.15 the classifier precision, determined by 10-fold cross-validation, is positively correlated with the random forest confidence value cut-off. In particular, at the cutoff value of 0.6 the classifier precision is 0.72, 0.9, 0.83, and 0.88 for SP, ME, obligate, and non-obligate classification, respectively. In other words, it achieves precision of over 0.8 for each classification problem, except for SP classification. However, since 21% of the SP/ME interactions are SP, a random SP classifier achieves a precision of 0.21, which means that our classifier is considerably better than a random classifier. Furthermore, the confidence value cutoff of 0.6 seems an acceptable trade off between precision and the number of classified interactions. Another reason to choose 0.6 as a cut-off value is that it guarantees the difference in confidence values between the opposing classes of at least 0.2. Note that confidence values are calculated by weighted majority voting. This means that at least 60% of the weighted random forest trees decided in favor of the chosen class and at most 40% for the opposing class, which implies that the classifier decision is based on a distinct majority.

For random forest confidence values >= 0.6 the total of 638 and 506 HeLa interactions as well as 1747 and 2620 iRefIndex interactions were classified as SP/obligate and SP/non-obligate, respectively. The total of 1010 and 6118 HeLa interactions were classified as ME/obligate and ME/non-obligate, respectively, while for the iRefIndex dataset the corresponding numbers were 1772 and 54580.

It was recently suggested that SP interactions are mostly permanent and ME inter-

actions are mostly transient [125]. As discussed above transient interactions are by definition non-obligate while permanent interactions are mostly obligate. In line with the results reported in [125] most of the ME interactions were classified as non-obligate both in the HeLa and iRefIndex datasets, presumably because proteins involved in ME interactions compete for the same binding side, which is only possible when the interactions are non-obligate. However, we found that 44% of the SP interactions in the HeLa dataset and 59% of the SP interactions in the iRefIndex were classified as non-obligate (compare 3.13, and 3.14). This result implies that a multimeric protein complex can either exist as a stable compound throughout its entire lifetime or it can dynamically form and dissolve during its lifetime. An example for a non-obligate multimeric protein complex are coat proteins involved in formation of molecular vesicles. The coat proteins associate together to form the coat of the molecular vesicle and upon delivering their payload they dissolve again from each other.

We also evaluated the classification results for iRefIndex interactions measured by different experimental methods, focusing on yeast two hybrid (Y2H) essay and tandem affinity purification (TAP). It was suggested that TAP has a preference for detecting obligate interactions while Y2H has no bias towards obligate or non-obligate interactions [166]. Indeed, as shown in Figure 3.16, interactions determined by TAP get classified as obligate three times more often that those measured by Y2H. At the same time the fraction of SP interactions increases by 40% from 0.10 in Y2H to 0.14 in TAP, in line with the previous observation that around half of the SP interactions are also obligate.

3.3.1 Protein complex analysis

We further applied our method to classify all intra- and inter-complex interactions in the CORUM and HeLa datasets. The overlap between different class combinations in terms of GO categories associated with them is very low, implying that each interaction type is intrinsic for a distinct set of cellular functions. As expected, most of the inter-complex interactions in each dataset (CORUM, HeLa) were classified as ME/non-obligate, indicating that protein complexes interact mostly transiently with each other.

With regard to intra-complex interactions we observed that small protein complexes possess high information content and thus tend to be enriched in just one interaction type (Figure 3.17). Larger protein complexes generally display increased diversity in terms of interaction types (except for complexes of size 8 in the HeLa dataset for which the sample size is very small), probably because they may contain functionally specialized subcomplexes, each with its own prevailing interaction type. For example, RNA polymerase II and the transcription factor TFIIH form an obligate/SP sub-compartment while TFEII, TFFII , and TFIIB are mostly involved in non-obligate/ME interactions, and the

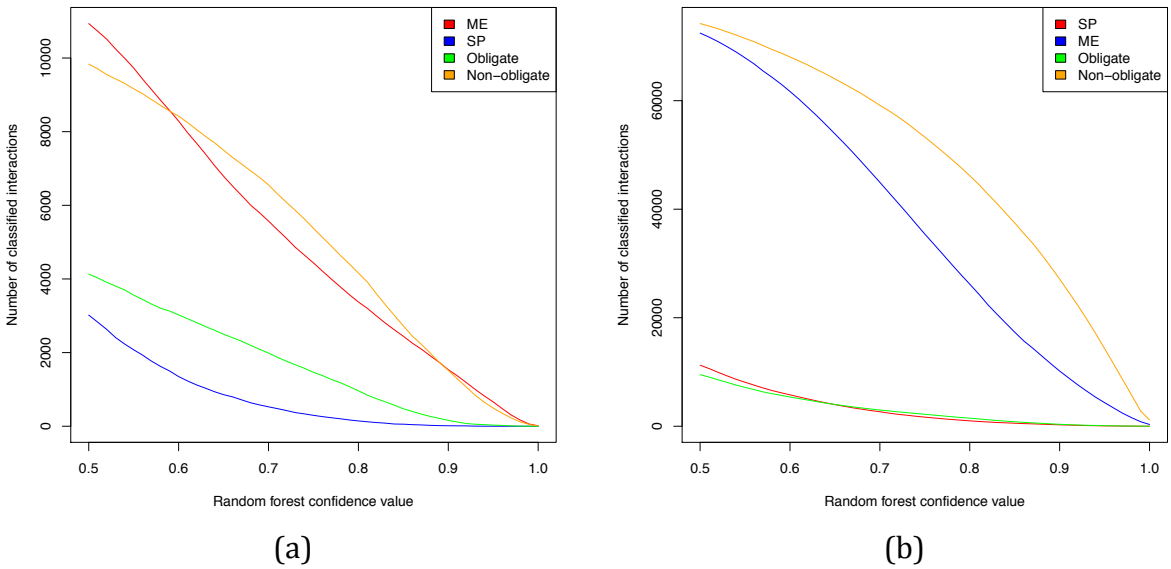
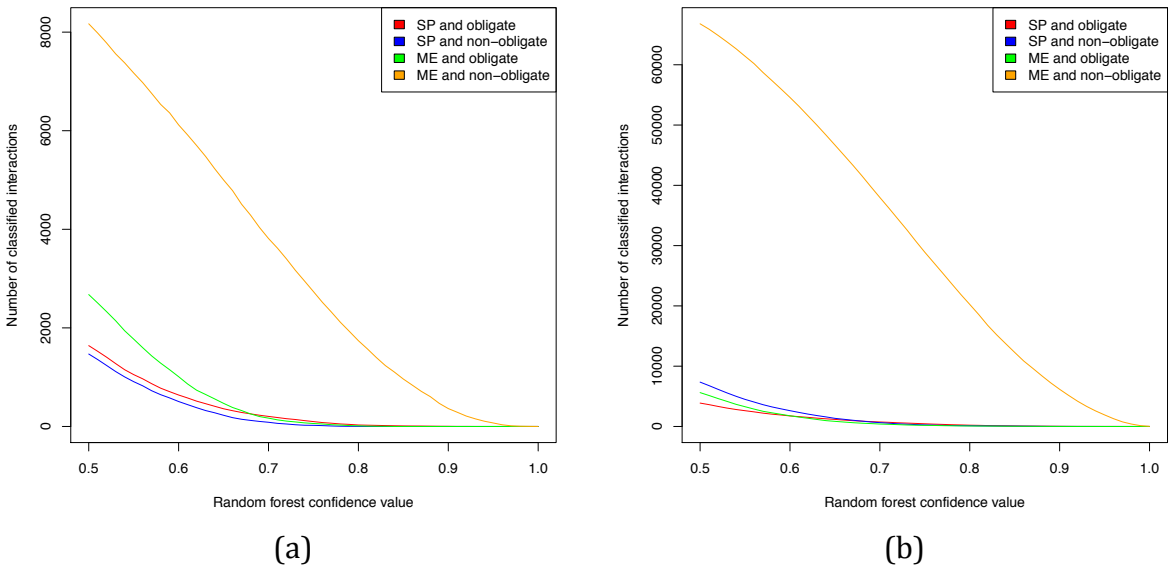


Figure 3.13: Number of classified interactions for each class for various random forest confidence cutofs in the HeLa dataset (a) and iRefIndex (b) dataset.

interactions between TFHII and the RNA polymerase II are also mostly non-obligate/ME (Figure 3.18).

Knowledge about interaction types can be instrumental for assessing the quality of protein complexes derived by computational methods. For example, the predicted mini-chromosome maintenance (MCM) complex (HeLa ID 587, Figure 3.19) consists of an obligate/SP part and a non-obligate/ME part. The obligate part exactly matches the CORUM MCM complex (CORUM ID 387), which is essential for DNA replication, initiation, and elongation in eukaryotic cells. The non-obligate/ME part is a novel addition to the MCM complex, which consist of the following proteins: amidophosphoribosyltransferase, RNA-binding protein 12B, splicing factor 3A subunit, and testis-specific serine kinase substrate. These proteins do not have any biological function associated with DNA replication, initiation, and elongation and it is probably safe to assume that they constitute false positive predictions added in the predicted HeLa complex to the manually verified CORUM complex.

We defined a protein cluster to be enriched in a given interaction type when it constituted at least 50% of the intra-complex interactions and plotted the fraction protein



(a) (b)

Figure 3.14: Number of classified interactions for all possible class combinations (SP and obligate, SP and non-obligate, ME and obligate, and ME and non-obligate) in the HeLa (a) and iRefIndex (b) dataset.

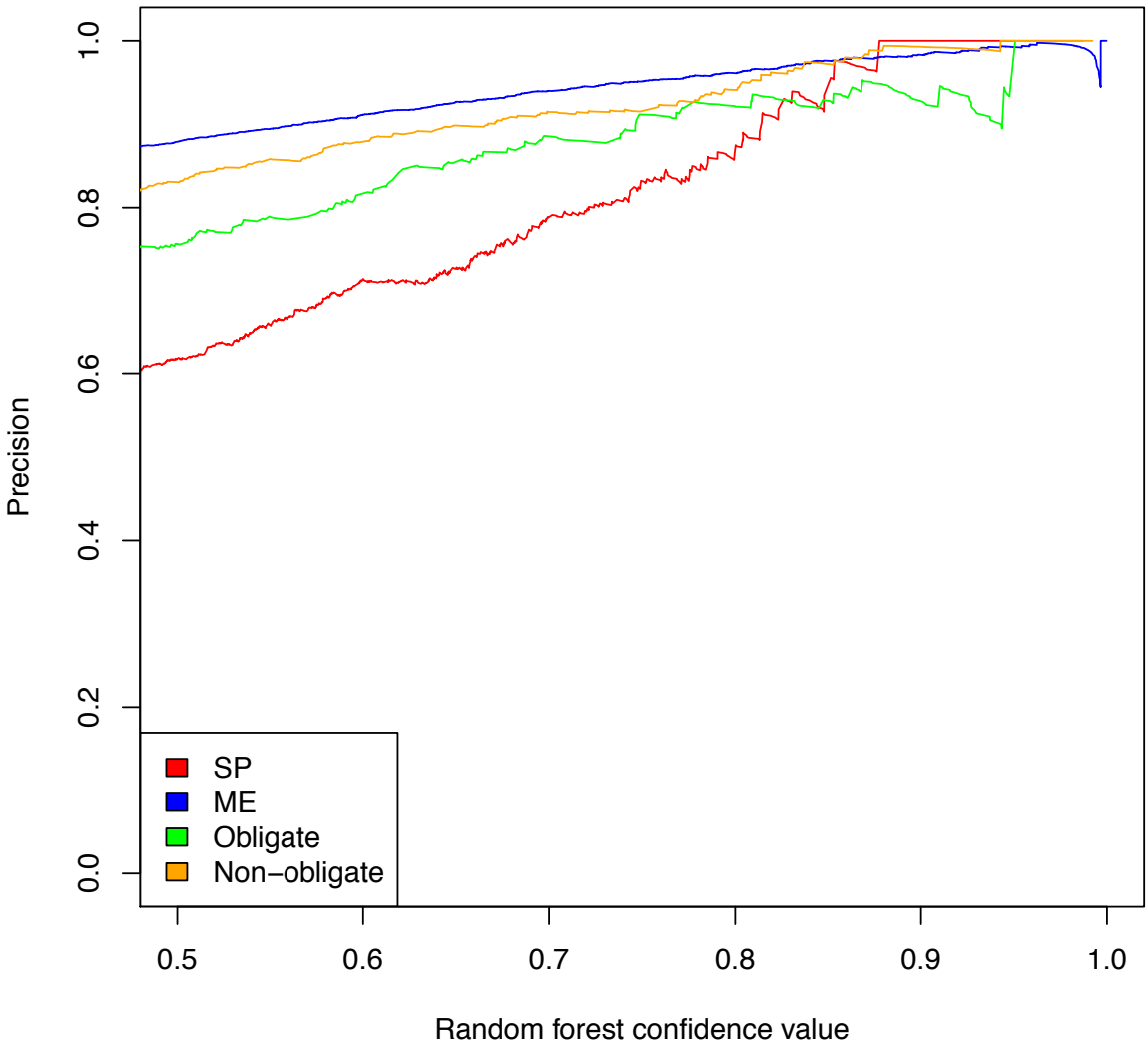


Figure 3.15: Dependence of the classifier precision on random forest confidence value cutoff in a 10-fold cross-validation.

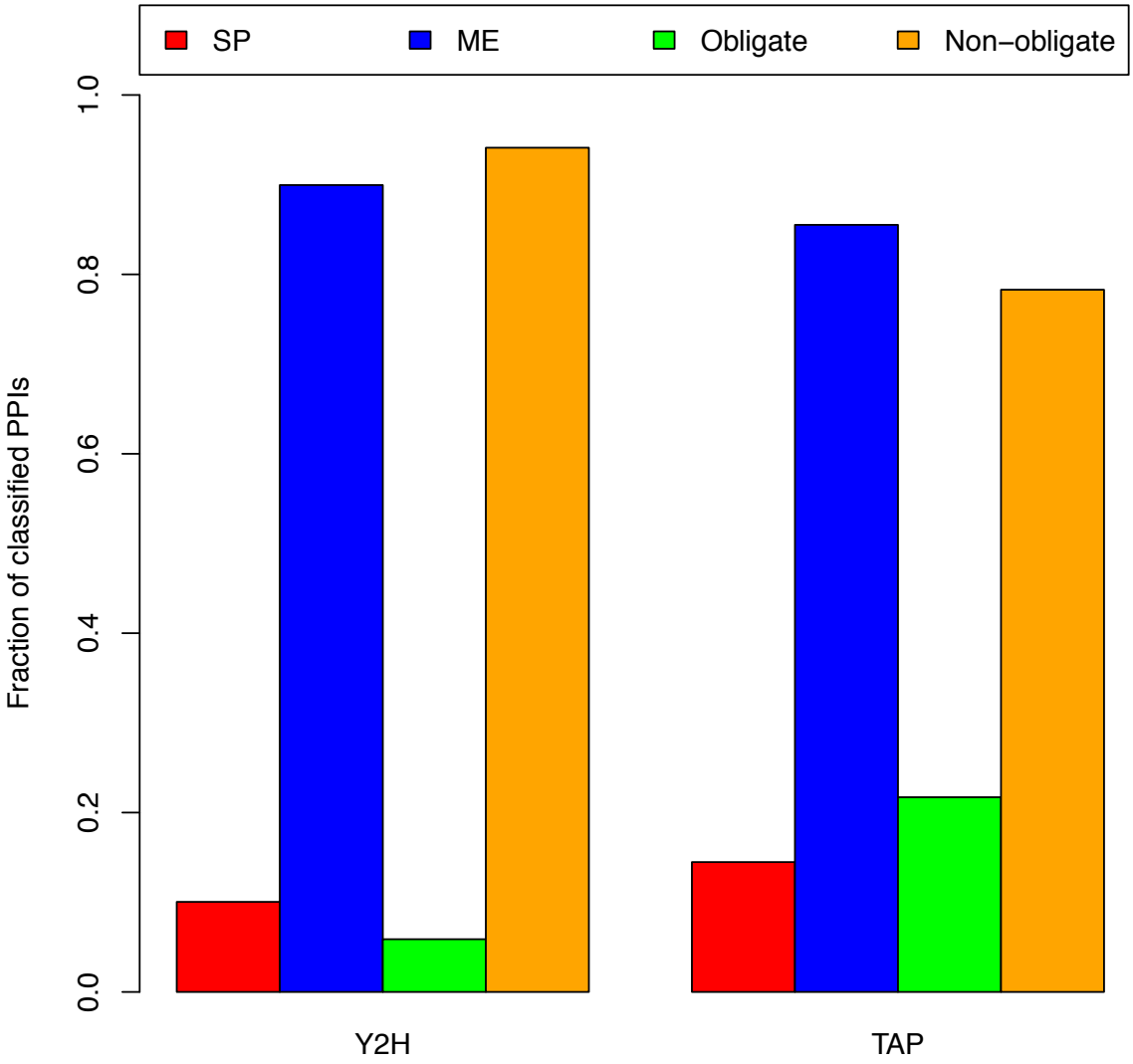


Figure 3.16: Class distributions for predicted protein interactions measured by the yeast two hybrid (Y2H) and tandem affinity purification (TAP) methods.

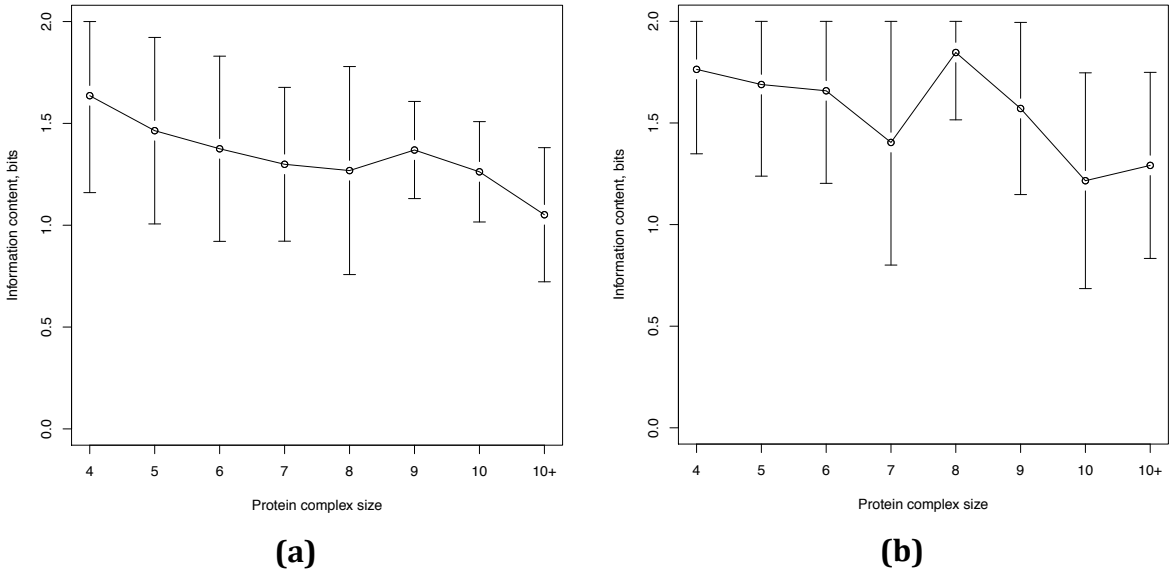


Figure 3.17: Information content vs protein complex size in the CORUM (a) and HeLa (b) datasets. Dots indicate the mean value of the information content and the error bars show its standard deviation.

complexes enriched in each interaction type (Figure 3.20). Both in the HeLa and in the CORUM datasets most of the protein complexes are enriched in ME/non-obligate interactions due to the fact that most of the ME interactions are non-obligate and the latter are frequently involved in intracellular signal transduction [159]. Correspondingly, ME/non-obligate interactions are enriched in GO terms associated with biological process regulation. Furthermore, around 50% of the protein complexes in the HeLa dataset are enriched in SP interactions whereas in the CORUM dataset only 25% of the complexes are SP-heavy. We speculate that the reason for this discrepancy lies in the somewhat different nature of these two datasets. The CORUM dataset used in this work was generated by overlaying multi-protein complexes described in the CORUM database with the binary interactions from the iRefIndex resource, while the HeLa dataset was derived by its authors by applying the CLusterOne method to a high confidence PPI network.

3.3.2 GO enrichment analysis

We applied GO enrichment analysis to explore the functional context of various types of interactions. As seen in Table 3.9 the overlap between different class combinations in terms of GO categories associated with them is very low, implying that each interaction type is intrinsic for a distinct set of cellular functions. The only deviation from this trend

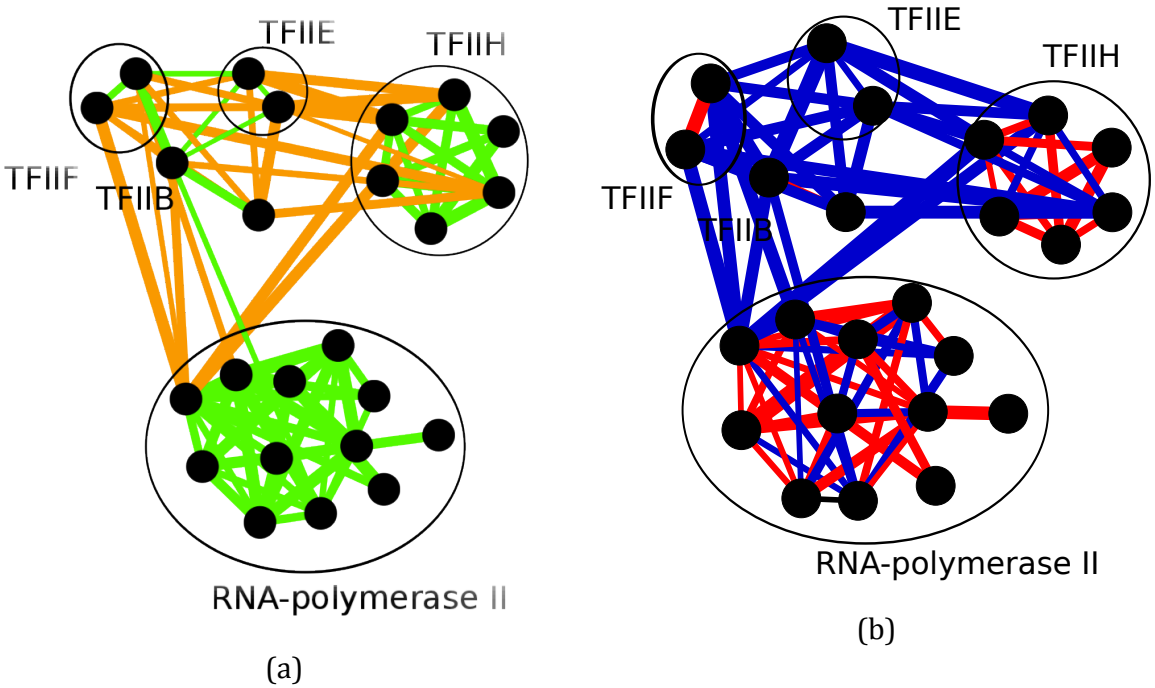


Figure 3.18: Protein interactions within the RNA polymerase II holoenzyme complex (CORUM ID: 103) classified as obligate (green) vs non-obligate (orange) (a) and SP (red) vs ME (blue) (b).

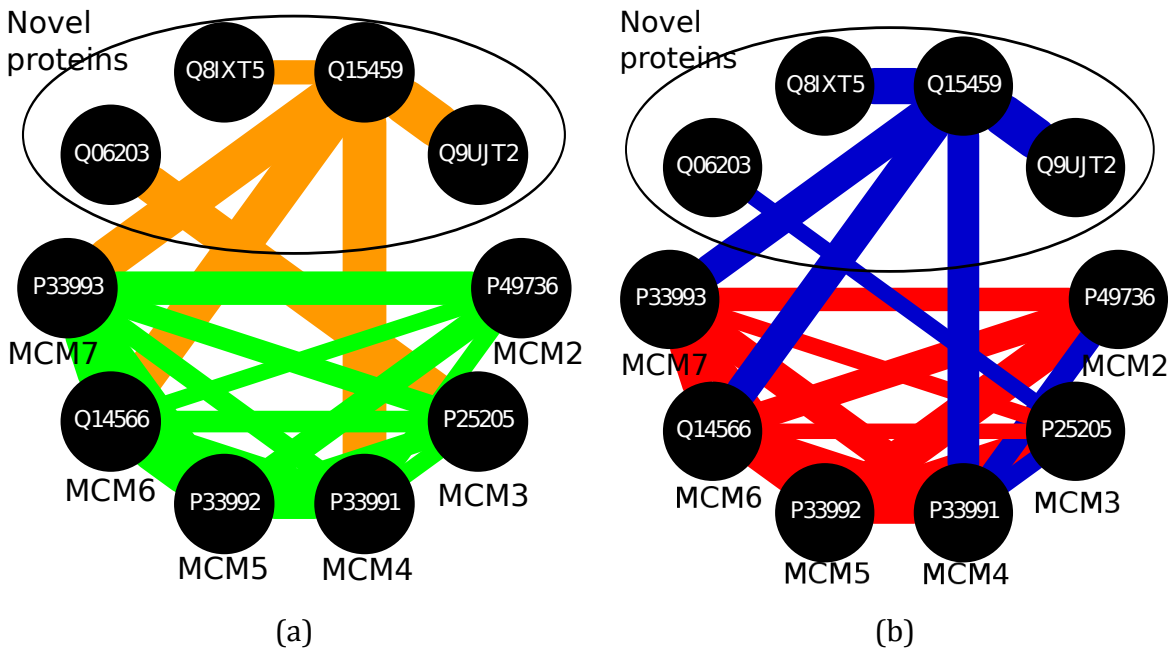


Figure 3.19: Protein interactions within the mini-chromosome maintenance (MCM) complex (HeLa ID: 587) classified as obligate (green) vs non-obligate (orange) and SP (red) vs ME (blue) (b). Uniprot accession numbers are shown for each protein. Additionally gene names are shown for members of the CORUM MCM complex.

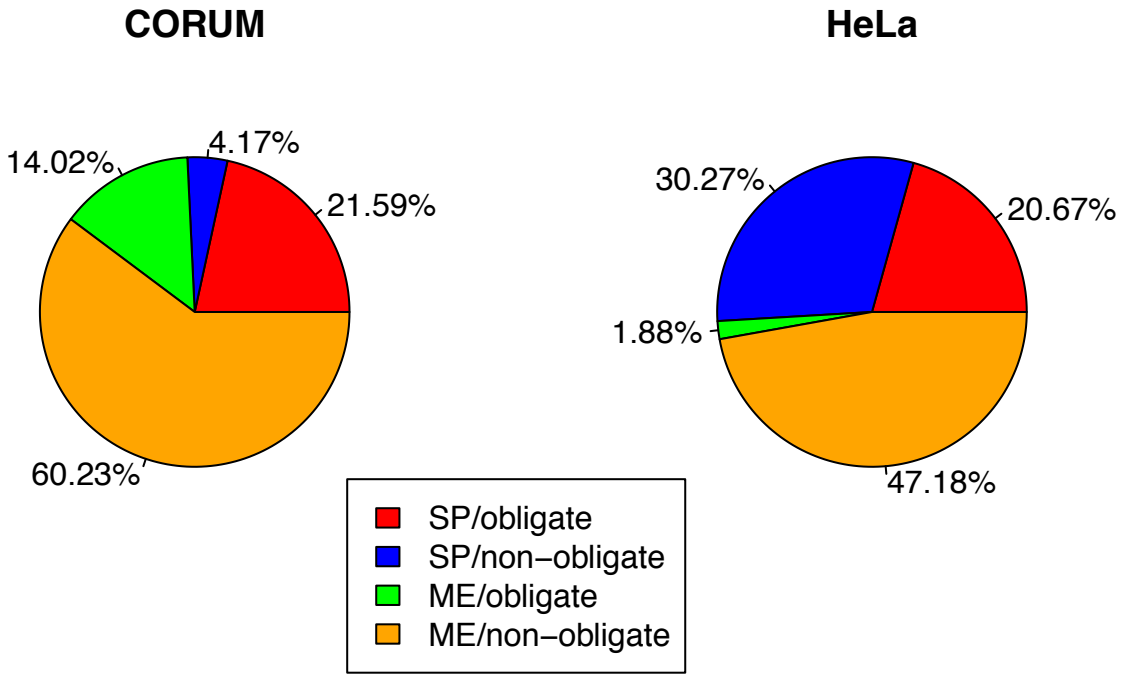


Figure 3.20: Fraction of enriched protein complexes in each dataset.

is the noticeable cross-talk between obligate/SP and obligate/ME interactions that share 52 GO terms, but only 8 of those actually describe molecular function while the remaining 44 shared GO terms refer to biological processes and cellular components. These two class combinations thus appear to share the same cellular location and to be involved in the same biological processes, yet to carry out distinctly different molecular functions. Table 3.10 shows a manually curated selection of the highest ranked (according to P-value) enriched GO terms. SP/obligate interactions are enriched in GO terms associated with nucleotide and nucleoside biosynthesis, their catabolism, as well as DNA replication and transcription. Proteins involved in these processes form stable multimeric complexes where they interact with their partners simultaneously. SP/non-obligate interactions frequently mediate cell-cell signaling. As for ME/obligate interactions, they are mostly associated with the GO terms describing complex subunit organization and lipid metabolic process and frequently occur in complexes with a ring shaped quaternary structure, such as the fatty acid synthase, he proteasome, and the U1 splicosome. ME/non-obligate interactions are enriched in GO terms describing the regulation of various biological processes and seem to play a key role in signal transduction.

Obligate and SP	58/107/273			
Obligate and ME	8/21/23	23/57/55		
Non-obligate and SP	3/0/1	0/0/0	8/15/34	
Non-obligate and ME	0/2/1	1/4/0	0/1/9	28/33/115
	Obligate and SP	Obligate and ME	Non-obligate and SP	Non-obligate and ME

Table 3.9: Number of enriched GO terms for each class combination (diagonal line) and number of overlapping GO terms for each pair of class combinations (non-diagonal entries). Each cell contains counts of molecular function, cellular component, and biological process GO terms.

	enriched GO terms
Obligate and SP	Nucleotide/nucleoside biosynthetic/catabolic processes
	DNA/RNA polymerase
	DNA replication
	RNA transcription
Non-obligate and SP	Cell adhesion
	Cell communication
	Locomotion
	Cell junction assembly
	Cell recognition
	Cell-cell signaling
	Generation of a signal involved in cell-cell signaling
	Cell projection organization
	Cell junction organization
Obligate and ME	Membrane lipid metabolic process
	Fatty acid elongation
	Lipid biosynthetic process
	Macromolecular complex subunit organization
	Ribonucleoprotein complex subunit organization
	Cellular macromolecular complex subunit organization
Non-obligate and ME	Positive/negative regulation of biosynthetic process
	Positive/negative regulation of metabolic process
	Positive/negative regulation of cellular process

Table 3.10: Manual non-redundant selection of the highest ranked enriched GO terms for each interaction type.

Chapter 4

Conclusion

In the last years, a large effort was made to improve methods for measuring physical protein protein interactions, which enables scientist to create complete interaction networks for several model organisms [222]. However, most of those methods only measure whether or not two protomers interact, and give no qualitative readout (i.e. binding affinity). Furthermore, methods for measuring binding affinity in protein interactions are in general not suitable to be conducted in a large scale manner [231].

The initial goal of this thesis was to create a method which is capable to quantitatively (i.e. the actual value) or qualitatively (i.e. weak/strong) predict the binding affinity of an interaction using only structure independent information. At first we used a small set of measured protein binding affinities [121] and overlayed it with the scores taken from the STRING [70] database. However, the overlap was insufficiently large enough for achieving any statistical significant results. In conclusion we changed the subject from predicting binding affinities to generally classifying protein interactions into respective interaction types such as obligate, non-obligate, SP, and ME.

Concerning protein interaction types, the most important paper was published in 2003 by Irene M. A. Nooren and Janet M. Thornton in which the terms transient, permanent, non-obligate, and obligate protein interface were defined [159]. The remaining research focused on finding differences based on sequence and structural between those interactions types [162]. It was shown that interfaces of non-obligate interactions tend to be smaller, less tightly packed, more polar, less conserved, and overall more similar to normal protein surfaces in terms of amino acid composition than those of obligate interactions [166].

In turn, those structural differences between obligate/non-obligate interactions were used to automatically classify protein complexes with known three-dimensional structure into various types based on physical, chemical, geometrical, and evolutionary properties

[17, 241, 164, 136, 183, 141, 23, 161, 151]. Most notably is the NOXclass classifier developed by Zhu et al. [241], since it is the only reported classifier which provides an online web server. NOXclass uses differences in interface area, amino acid composition, shape complementarity, and evolutionary conservation to classify protein interface into obligate or non-obligate.

The second type of interactions investigated in this thesis are simultaneously possible and mutually exclusive interactions [125]. So far there is no reported automatic method for predicting those interaction types. A recent study showed that, protein in SP interactions tend to be functionally more similar to each other compared to protein undergoing ME binding. Furthermore, they showed that proteins that bind multiple partners at the same time tend to under go larger conformational changes upon binding [14].

In this thesis, we created PiType the worlds first structure independent classifier for protein interaction types. We extensively evaluated our classifier using a variety of metrics and tests. First we used machine learning based metrics such as n-fold cross-validation, while the second part consist of a biological evaluation based on a large scale application of PiType.

We created a naïve classifier, which allows us to compare our classifier to some baseline performance. This was necessary since no other method for comparison was available. We observed that PiType works considerably better than the naïve classifier, and thus justifies the creation of PiType.

We also applied 10-fold cross-validation for both obligate/non-obligate and SP/ME classification. We observed that the classifier achieves an auROC of at least 80% and a F-measure close to 80%, which is comparable with that of the structure-based classifiers. Furthermore, we observed only a small deviation between evaluation metrics taken from 10-fold cross-validation and nested cross-fold validation. This suggest that the PiType is a robust and not overfitted classifier, which means that its performance should be stable for new, or different data sets.

At last we performed a cross-species evaluation, in which we trained on one species set and evaluated the classifier on a different one (e.g. trained on Yeast and evaluated on Human data). The resulting auROC values showed that the classifier performance is consistent across different species. It is almost the same for obligate/non-obligate classification and around $1\% - 5\%$ are lost in SP/ME classification. Thus, the results suggests that both the obligate/non-obligate and SP/ME classifiers can be used the classify interaction data from multiple species, which justifies our later inclusion of the STRING database.

Feature selection was executed in order to find the best features for distinguishing the two classification problems: obligate/non-obligate, and SP/ME. In order to determine these features we grouped them into logical groups: network based features, sequence

based features, and functional similarity.

Concerning network based features, we showed that they play almost no importance for differentiating obligate from non-obligate interactions. Although, we did found out that obligate interactions have a preference for having a highly connected local topology such as cliques, or network clusters, while non-obligate interactions are enriched in orbits describing hub like proteins. This is, according to the observation that obligate interactions occur mostly in stable protein complexes, whereas non-obligate interactions are associated with signaling pathways [166, 162].
We observed network based features differ greatly between SP and ME interactions. We showed that the most important features to distinct SP interaction from ME is the edge degree. Most likely, the reason for this is the physical limit to how many interaction partners can simultaneously bind to a protein. Furthermore, we also observed significant differences in local topology. SP interactions are enriched in orbits describing sparsely connected topologies, which might also be related to the physical limits of binding multiple partners simultaneously. As for ME interactions, they are enriched in orbits describing bottlenecks, which implies they serve as key connectors.

Furthermore, we found significant results for both sequence based features: short linear eukaryotic motifs, and the number of predicted disordered regions. We observed that non-obligate interactions have considerably more disordered regions than obligate interactions, which is in agreement with the latest scientific consensus [194]. Furthermore, non-obligate interactions are enriched in short linear eukaryotic motifs (ELMs), which is in line with the notion that that ELM primarily mediate weak transient interactions occurring in signaling [198].
As for proteins which take part in SP interactions, we observed that they have a tendency to be more disordered than those in mutually exclusive interactions, because simultaneously possible interactors undergo stronger conformational changes upon binding with their partners than mutually exclusive interactors [14].

Last but not least, we also investigated the functional similarity of representative protein pairs for each interaction type. We showed that interaction undergoing in SP, or obligate interactions are considerably more functional similar to each other than protein which interact via a ME or non-obligate interaction.

Additional to the machine learning based analysis, we also conducted a large scale biological investigation of the classifier performance. For this we applied the final PiType pipeline to classify 13978 HeLa [98] and 83788 iRefIndex [177] protein interactions. In the next step we overlaid those classified interaction network with protein complex information taken from the HeLa data set, and the CORUM database [184].

First, we needed to identify the optimal predictor confidence value cut-off, in order to create a confident and trustworthy classified interaction network. For this we com-

pared recall and precision values for various confidence values, and determined that the optimal cut-off is 0.6, since it allows a good trade off between the predictors precision and fraction of classified protein interactions in both the HeLa and iRefIndex network.

In the next step we took all interaction in the iRefIndex which were measured ether with yeast two hybrid (Y2H) essay or with tandem affinity purification (TAP). We decided to use these two methods since it was reported that TAP has a preference for detecting obligate interactions while Y2H has no bias towards obligate or non-obligate interactions [166]. As expected, we observed that most of the interactions from iRefIndex which were measured by TAP were classified as obligate, and no significant change was observed for Y2H interactions, which serves as an additional validation for our classifier.

In the last analysis we applied PiType to classify all intra- and inter-complex interactions in the CORUM and HeLa datasets. We observed that most of the inter protein complexes are ME and non-obligate, which suggest that protein complexes interact mostly transiently with each other. Concerning intra protein complex interactions we showed that smaller protein complexes tend to be enriched in one interaction type, whereas larger protein complexes may contain a diversity of interaction types, since larger protein complexes can contain many functionally specialized sub-complexes. Furthermore, GO enrichment analysis showed that each interaction type is intrinsic for a distinct set of cellular functions.

The second part of my thesis focused on improving the previously established PiType pipeline, leading to the development of PiType 2.0, an updated version of PiType. The Major enhancement is the inclusion of the STRING database as a reference network, which greatly increase the number of classifiable interactions and supported species. However we had to reevaluate PiType, since the STRING database contains functional associations, whereas the previously used iRefIndex database stored only direct physical interactions. Thus, the classifier performance was evaluated by conducting 10-fold cross-validation on our SP/ME and obligate/non-obligate training data with various STRING score cut-offs. Overall, we showed that the optimal cut-off is 950, because at this cut-off we achieved maximal classifier performance, that is also similar to the original PiType. The second major improvement we have done to the PiType pipeline was to create a freely accessible web interface. Similar to the NOXclass classifier we wanted to create a simple stream-lined interface to ensure an user friendly usage. The user has to submit a list of physical interacting protein pairs, and then can optionally select which feature should be applied and which target species should be used. The server accepts identifiers from several source database, and maps the supplied identifiers onto Uniprot IDs of the supplied species. In cases where no target species were supplied, the server creates a list of species sorted by the number of mapped proteins from which the user can select the desired targeted species.

Chapter 5

Outlook

In this chapter I will elaborate possible further projects. Possible strategies for the continuation of the Pitype pipeline can be split into two general approaches: a) improvement of the prediction pipeline, and b) possible applications for various biological problems. There are several ways in which PiType can be modified. One crucial improvement would be to predict protein interaction types using only sequence based features. This can be achieved for obligate/non-obligate classification, since it does not rely heavily on network based features. However, the greatest improvement would be the inclusion of predicting binding regions/interfaces and to classify them into obligate/non-obligate and into SP/ME. Another great improvement for SP/ME interactions would be to no longer consider single PPIs, but rather classify pairs of protein interactions.

As for applying PiType for biological problems, we are interested in researching how the interaction types relate to paralogous proteins. Precisely, we are interested if interaction types can be used to distinguish novo-functionalization from sub-functionalization. The second project would be to analyze host/pathogen interactions. The concrete question would be: What are the differences between transient and permanent host-pathogen interactions?

Bibliography

[1] Mark D Adams, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, Richard F Galle, et al. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[2] Sam Ansari and Volkhard Helms. Statistical analysis of predominantly transient protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 61(2):344–355, 2005.

[3] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

[4] Heidwolf Arnold and Dirk Pette. Binding of aldolase and triosephosphate dehydrogenase to f-actin and modification of catalytic properties of aldolase. *European Journal of Biochemistry*, 15(2):360–366, 1970.

[5] Ken Arnold, James Gosling, and David Holmes. *The Java programming language*, volume 2. Addison-wesley Reading, 1996.

[6] Ami Aronheim, Ebrahim Zandi, Hanjo Hennemann, Stephen J Elledge, and Michael Karin. Isolation of an ap-1 repressor by a novel method for detecting protein-protein interactions. *Molecular and cellular biology*, 17(6):3094–3102, 1997.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[8] Paul L Bartel, Jennifer A Roecklein, Dhruba SenGupta, and Stanley Fields. A protein linkage map of escherichia coli bacteriophage t7. *Nature genetics*, 12(1):72–77, 1996.

[9] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[10] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N Robinson. Ontologizer 2.0 — a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, 2008.

[11] Tim Beißbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[12] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and David L Wheeler. Genbank. *Nucleic acids research*, 36(suppl 1):D25–D30, 2008.

[13] Frances C Bernstein, Thomas F Koetzle, Graheme JB Williams, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank. *European Journal of Biochemistry*, 80(2):319–324, 1977.

[14] Nitin Bhardwaj, Alexej Abyzov, Declan Clarke, Chong Shou, and Mark B Gerstein. Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein science*, 20(10):1745–1754, 2011.

[15] Nitin Bhardwaj and Hui Lu. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

[16] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[17] Peter Block, Juri Paern, Eyke Huellermeier, Paul Sanschagrin, Christoph A Sotriffer, and Gerhard Klebe. Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *PROTEINS: Structure, Function, and Bioinformatics*, 65(3):607–622, 2006.

[18] Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

[19] David C Boisvert, Jimin Wang, Zbyszek Otwinowski, Arthur L Norwich, and Paul B Sigler. The 2.4 å crystal structure of the bacterial chaperonin groel complexed with atpγs. *Nature Structural & Molecular Biology*, 3(2):170–177, 1996.

[20] Jaume Bonet, Gianluigi Caltabiano, Abdul Kareem Khan, Michael A Johnston, Carles Corbí, Àlex Gómez, Xavier Rovira, Joan Teyra, and Jordi Villà-Freixa. The role of residue stability in transient protein–protein interactions involved in enzymatic phosphate hydrolysis. a computational study. *Proteins: Structure, Function, and Bioinformatics*, 63(1):65–77, 2006.

[21] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. *Annual review of biophysics and biomolecular structure*, 30(1):173–189, 2001.

[22] David Botstein and Gerald R Fink. Yeast: an experimental organism for modern biology. *Science*, 240(4858):1439–1443, 1988.

[23] James R Bradford and David R Westhead. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494, 2005.

[24] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[25] Pascal Braun and Anne-Claude Gingras. History of protein–protein interactions: From egg-white to complex networks. *Proteomics*, 12(10):1478–1498, 2012.

[26] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[27] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[28] James A Brown, Gavin Sherlock, Chad L Myers, Nicola M Burrows, Changchun Deng, H Irene Wu, Kelly E McCann, Olga G Troyanskaya, and J Martin Brown. Global analysis of gene function in yeast by quantitative phenotypic profiling. *Molecular systems biology*, 2(1), 2006.

[29] Axel T Brunger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, et al. Crystallography & nmr system: a new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):905–921, 1998.

[30] Rafael Bruschweiler, Xiubei Liao, and Peter E Wright. Long-range motional restrictions in a multidomain zinc-finger protein from anisotropic tumbling. *Science*, 268(5212):886–889, 1995.

[31] Arnaud Ceol, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. Mint, the molecular interaction database: 2009 update. *Nucleic acids research*, 38(suppl 1):D532–D539, 2010.

[32] Pinak Chakrabarti and Joel Janin. Dissecting protein–protein recognition sites. *Proteins: Structure, Function, and Bioinformatics*, 47(3):334–343, 2002.

[33] Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, et al. The biogrid interaction database: 2013 update. *Nucleic acids research*, 41(D1):D816–D823, 2013.

[34] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 2011.

[35] Nitesh V Chawla, David A Cieslak, Lawrence O Hall, and Ajay Joshi. Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, 17(2):225–252, 2008.

[36] Asif T Chinwalla, Lisa L Cook, Kimberly D Delehaunty, Ginger A Fewell, Lucinda A Fulton, Robert S Fulton, Tina A Graves, LaDeana W Hillier, Elaine R Mardis, John D McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[37] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366–2382, 2007.

[38] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[39] The Uniprot consortium. The uniprot documentation, 2014.

[40] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198, 1999.

[41] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[42] FHC Crick and LE Orgel. The theory of inter-allelic complementation. *Journal of molecular biology*, 8(1):161–165, 1964.

[43] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 2006.

[44] Pedro Cuatrecasas. Protein purification by affinity chromatography derivatizations of agarose and polyacrylamide beads. *Journal of Biological Chemistry*, 245(12):3059–3065, 1970.

[45] Norman E Davey, Gilles Travé, and Toby J Gibson. How viruses hijack cell regulation. *Trends in biochemical sciences*, 36(3):159–169, 2011.

[46] Norman E Davey, Kim Van Roey, Robert J Weatheritt, Grischa Toedt, Bora Uyar, Brigitte Altenberg, Aidan Budd, Francesca Diella, Holger Dinkel, and Toby J Gibson. Attributes of short linear motifs. *Molecular BioSystems*, 8(1):268–281, 2012.

[47] Warren L DeLano. Unraveling hot spots in binding interfaces: progress and challenges. *Current opinion in structural biology*, 12(1):14–20, 2002.

[48] Sucharita Dey, Arumay Pal, Pinak Chakrabarti, and Joël Janin. The subunit interfaces of weakly associated homodimeric proteins. *Journal of molecular biology*, 398(1):146–160, 2010.

[49] Francesca Diella, Niall Haslam, Claudia Chica, Aidan Budd, Sushama Michael, Nigel P Brown, Gilles Travé, and Toby J Gibson. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*, 13:6580–6603, 2008.

[50] Holger Dinkel, Sushama Michael, Robert J Weatheritt, Norman E Davey, Kim Van Roey, Brigitte Altenberg, Grischa Toedt, Bora Uyar, Markus Seiler, Aidan Budd, et al. Elm—the database of eukaryotic linear motifs. *Nucleic acids research*, 40(D1):D242–D251, 2012.

[51] Holger Dinkel, Kim Van Roey, Sushama Michael, Norman E Davey, Robert J Weatheritt, Diana Born, Tobias Speck, Daniel Krüger, Gleb Grebnev, Marta Kubań, et al. The eukaryotic linear motif resource elm: 10 years and counting. *Nucleic acids research*, page gkt1047, 2013.

[52] Scott J Dixon, Michael Costanzo, Anastasia Baryshnikova, Brenda Andrews, and Charles Boone. Systematic mapping of genetic interaction networks. *Annual review of genetics*, 43:601–625, 2009.

[53] Zsuzsanna Dosztányi, Bálint Mészáros, and István Simon. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Briefings in bioinformatics*, page bbp061, 2009.

[54] Paul DuBois. *MySQL (Developer's Library)*. Sams, 2005.

[55] A Keith Dunker, Zoran Obradovic, Pedro Romero, Ethan C Garner, Celeste J Brown, et al. Intrinsic protein disorder in complete genomes. *Genome Informatics Series*, pages 161–171, 2000.

[56] Ruth Dunn, Frank Dudbridge, and Christopher M Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC bioinformatics*, 6(1):39, 2005.

[57] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology*, 12(1):54–60, 2002.

[58] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology*, 6(3):197–208, 2005.

[59] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, 2009.

[60] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.

[61] Santiago F Elena and Richard E Lenski. Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, 390(6658):395–398, 1997.

[62] Sarah J Fashena, Ilya Serebriiskii, and Erica A Golemis. The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits. *Gene*, 250(1):1–14, 2000.

[63] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein protein interactions. 1989.

[64] BC Finzel, PC Weber, KD Hardman, and FR Salemme. Structure of ferricytochrome¡ i¿ c¡/i¿' from¡ i¿ rhodospirillum molischianum¡/i¿ at 1.67 åresolution. *Journal of molecular biology*, 186(3):627–643, 1985.

[65] Gary William Flake, Steve Lawrence, C Lee Giles, and Frans M Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–70, 2002.

[66] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, 41(D1):D48–D55, 2013.

[67] M Florkin. Discovery of pepsin by theodor schwann. *Revue médicale de Liège*, 12(5):139, 1957.

[68] CA Floudas, HK Fung, SR McAllister, M Mönnigmann, and R Rajgaria. Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, 61(3):966–988, 2006.

[69] James H Fowler. Legislative cosponsorship networks in the us house and senate. *Social Networks*, 28(4):454–465, 2006.

[70] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, et al. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

[71] Hunter B Fraser, Aaron E Hirsh, Dennis P Wall, and Michael B Eisen. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9033–9038, 2004.

[72] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[73] Alberto Freitas, Altamiro Costa-Pereira, and Pavel Brazdil. Cost-sensitive decision trees applied to medical data. In *Data Warehousing and Knowledge Discovery*, pages 303–312. Springer, 2007.

[74] Haian Fu. *Protein-protein interactions: methods and applications*, volume 261. Springer, 2004.

[75] Monika Fuxreiter, Peter Tompa, and István Simon. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956, 2007.

[76] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):463–484, 2012.

[77] Charles A Galea, Amanda Nourse, Yuefeng Wang, Sivashankar G Sivakolundu, William T Heller, and Richard W Kriwacki. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27¡ sup¿ kip1¡/sup¿. *Journal of molecular biology*, 376(3):827–838, 2008.

[78] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[79] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, pages 225–234. ACM, 1998.

[80] Loic Giot, Joel S Bader, C Brouwer, Amitabha Chaudhuri, Bing Kuang, Y Li, YL Hao, CE Ooi, Brian Godwin, E Vitols, et al. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[81] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[82] Florian Goebels and Dmitrij Frishman. Prediction of protein interaction types based on sequence and network features. *BMC Systems Biology*, 7(6):1–18, 2013.

[83] André Goffeau, BG Barrell, Howard Bussey, RW Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, JD Hoheisel, Cr Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[84] Ian S Graham. *The HTML sourcebook*. John Wiley & Sons, Inc., 1995.

[85] Andrei Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic acids research*, 29(17):3513–3519, 2001.

[86] Jonathan L Gross and Jay Yellen. *Handbook of graph theory*. CRC press, 2003.

[87] Steffen Grossmann, Sebastian Bauer, Peter N Robinson, and Martin Vingron. Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. *Bioinformatics*, 23(22):3024–3031, 2007.

[88] Mainak Guharoy and Pinak Chakrabarti. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15447–15452, 2005.

[89] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[90] Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–806, 2005.

[91] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[92] Mark A Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[93] Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, Frederick P Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.

[94] James A Hanley, Barbara J McNeil, et al. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[95] Robert M Hanson, Jaime Prilusky, Zhou Renjian, Takanori Nakane, and Joel L Sussman. Jsmol and the next-generation web-based representation of 3d molecular structure as applied to proteopedia. *Israel Journal of Chemistry*, 53(3-4):207–216, 2013.

[96] Niall J Haslam and Denis C Shields. Peptide-binding domains: are limp handshakes safest? *Science signaling*, 5(243):pe40, 2012.

[97] Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003.

[98] Pierre C Havugimana, G Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L Turinsky, Zhihua Li, Peggy I Wang, Daniel R Boutz, Vincent Fong, Sadhna Phanse, et al. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, 2012.

[99] SG Hedin. Trypsin and antitrypsin. *Biochemical Journal*, 1(10):474, 1906.

[100] Angel Herraez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.

[101] John H Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[102] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[103] Simon J Hubbard and Janet M Thornton. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1), 1993.

[104] Lukas A Huber. Is proteomics heading in the wrong direction? *Nature Reviews Molecular Cell Biology*, 4(1):74–80, 2003.

[105] Martijn Huynen, Berend Snel, Warren Lathe, and Peer Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, 10(8):1204–1210, 2000.

[106] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradović, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology*, 323(3):573–584, 2002.

[107] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[108] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[109] Joël Janin, Ranjit P Bahadur, and Pinak Chakrabarti. Protein–protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(02):133–180, 2008.

[110] Ronald Jansen, Dov Greenbaum, and Mark Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome research*, 12(1):37–46, 2002.

[111] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. String 8 a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(suppl 1):D412–D416, 2009.

[112] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[113] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[114] Susan Jones and Janet M Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272(1):121–132, 1997.

[115] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[116] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, 2005(2):96–103, 2005.

[117] Anna k. Two hybrid assay, 2007.

[118] William G Kaelin. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698, 2005.

[119] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[120] Carola Kanz, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Alastair Baldwin, Kirsty Bates, Paul Browne, Alexandra van den Broek, Matias Castro, Guy Cochrane, et al. The embl nucleotide sequence database. *Nucleic Acids Research*, 33(suppl 1):D29–D33, 2005.

[121] Panagiotis L Kastritis, Iain H Moal, Howook Hwang, Zhiping Weng, Paul A Bates, Alexandre MJJ Bonvin, and Joël Janin. A structure-based benchmark for protein–protein binding affinity. *Protein Science*, 20(3):482–491, 2011.

[122] Tohru Kataoka, Scott Powers, Scott Cameron, Ottavio Fasano, Mitchell Goldfarb, James Broach, and Michael Wigler. Functional homology of mammalian and yeast¡ i¿ ras¡/i¿ genes. *Cell*, 40(1):19–26, 1985.

[123] Masaru Katoh. Wnt/pcp signaling pathway and human cancer (review). *Oncology reports*, 14(6):1583–1588, 2005.

[124] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. The intact molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846, 2012.

[125] Philip M Kim, Long J Lu, Yu Xia, and Mark B Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941, 2006.

[126] Stuart K Kim, Jim Lund, Moni Kiraly, Kyle Duke, Min Jiang, Joshua M Stuart, Andreas Eizinger, Brian N Wylie, and George S Davidson. A gene expression map for caenorhabditis elegans. *Science*, 293(5537):2087–2092, 2001.

[127] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[128] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P Tikuisis, et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[129] W Kuehne. Ueber das trypsin (enzyme des pankreas). *Heidelberg Nat Med Ver*, 1:194–198, 1876.

[130] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer networks*, 31(11):1481–1493, 1999.

[131] John E Ladbury, Mark A Lemmon, Min Zhou, Jeremy Green, Martyn C Botfield, and Joseph Schlessinger. Measurement of the binding of tyrosyl phosphopeptides to sh2 domains: a reappraisal. *Proceedings of the National Academy of Sciences*, 92(8):3199–3203, 1995.

[132] Roman A Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5):323–330, 1995.

[133] Jae Woon Lee and Soo-Kyung Lee. Mammalian two-hybrid assay for detecting protein-protein interactions in vivo. In *Protein-Protein Interactions*, pages 327–336. Springer, 2004.

[134] Ben Lehner, Catriona Crombie, Julia Tischler, Angelo Fortunato, and Andrew G Fraser. Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature genetics*, 38(8):896–903, 2006.

[135] Siming Li, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J Han, Alban Chesneau, Tong Hao, et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

[136] Qian Liu and Jinyan Li. Propensity vectors of low-asa residue pairs in the distinction of protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 78(3):589–602, 2010.

[137] Rong Liu, Wenchao Jiang, and Yanhong Zhou. Identifying protein–protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. *Amino acids*, 38(1):263–270, 2010.

[138] Yi-Hung Liu and Yen-Ting Chen. Total margin based adaptive fuzzy support vector machines for multiview face recognition. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1704–1711. IEEE, 2005.

[139] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, 100(10):5772–5777, 2003.

[140] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl 1):D54–D58, 2005.

[141] Mina Maleki, Md Mominul Aziz, and Luis Rueda. Analysis of obligate and non-obligate complexes using desolvation energies in domain-domain interactions. In *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics*, page 2. ACM, 2011.

[142] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[143] Elaine R Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[144] Ruschhaupt Markus, Huber Wolfgang, Poustka Annemarie, and Mansmann Ulrich. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–26, 2004.

[145] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2):427–436, 2008.

[146] Brendan D McKay. Nauty user's guide (version 2.4). *Computer Science Dept., Australian National University*, 2007.

[147] Lidio MC Meireles, Alexander S Dömling, and Carlos J Camacho. Anchor: a web server and database for analysis of protein–protein interaction binding pockets for drug discovery. *Nucleic acids research*, 38(suppl 2):W407–W411, 2010.

[148] Leonor Michaelis and Maud L Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[149] Tijana Milenkoviæ and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257, 2008.

[150] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[151] Julian Mintseris and Zhiping Weng. Atomic contact vectors in protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 53(3):629–639, 2003.

[152] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10930–10935, 2005.

[153] Satoru Miyazaki, Hideaki Sugawara, Kazuho Ikeo, Takashi Gojobori, and Yoshio Tateno. Ddbj in the stream of various biological data. *Nucleic Acids Research*, 32(suppl 1):D31–D34, 2004.

[154] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, 12(1):88–118, 1965.

[155] John Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3):285–289, 2005.

[156] Victor Neduva and Robert B Russell. Linear motifs: evolutionary interaction switches. *FEBS letters*, 579(15):3342–3345, 2005.

[157] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.

[158] Hani Neuvirth, Ran Raz, and Gideon Schreiber. Promate: a structure based prediction program to identify the location of protein–protein binding sites. *Journal of molecular biology*, 338(1):181–199, 2004.

[159] Irene Nooren and Janet M Thornton. Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492, 2003.

[160] Irene Nooren and Janet M Thornton. Structural characterisation and functional significance of transient protein–protein interactions. *Journal of molecular biology*, 325(5):991–1018, 2003.

[161] Yanay Ofran and Burkhard Rost. Analysing six types of protein–protein interfaces. *Journal of molecular biology*, 325(2):377–387, 2003.

[162] Saliha Ece Acuner Ozbabacan, Hatice Billur Engin, Attila Gursoy, and Ozlem Keskin. Transient protein–protein interactions. *Protein Engineering Design and Selection*, 24(9):635–648, 2011.

[163] Savas Parastatidis, Jim Webber, Simon Woodman, Dean Kuo, and Paul Greenfield. An introduction to the soap service description language. *School of Computing Science, University of Newcastle, Newcastle upon Tyne CS-TR-898*, 2005.

[164] Sung H Park, José A Reyes, David R Gilbert, Ji W Kim, and Sangsoo Kim. Prediction of protein-protein interaction types using association rule based classification. *BMC bioinformatics*, 10(1):36, 2009.

[165] Ashwini Patil, Kengo Kinoshita, and Haruki Nakamura. Hub promiscuity in protein-protein interaction networks. *International journal of molecular sciences*, 11(4):1930–1943, 2010.

[166] James R Perkins, Ilhem Diboun, Benoit H Dessailly, Jon G Lees, and Christine Orengo. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243, 2010.

[167] David Perrett. From protein to the beginnings of clinical proteomics. *Proteomics-Clinical Applications*, 1(8):720–738, 2007.

[168] Thomas D Pollard, Robert R Weihing, and MR Adelman. Actin and myosin and cell movemen. *Critical Reviews in Biochemistry and Molecular Biology*, 2(1):1–65, 1974.

[169] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database 2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[170] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. Refseq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763, 2014.

[171] Oscar Puig, Friederike Caspary, Guillaume Rigaut, Berthold Rutz, Emmanuelle Bouveret, Elisabeth Bragado-Nilsson, Matthias Wilm, and Bertrand Séraphin. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

[172] Amit Puniyani, Uri Liberman, and Marcus W Feldman. On the meaning of non-epistatic selection. *Theoretical population biology*, 66(4):317–321, 2004.

[173] Pål Puntervoll, Rune Linding, Christine Gemünd, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David MA Martin, Gabriele Ausiello, Barbara Brannetti, Anna Costantini, et al. Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic acids research*, 31(13):3625–3630, 2003.

[174] January Python. Python programming language. *Python (programming language) 1 CPython 13 Python Software Foundation 15*, page 1, 2009.

[175] J. Ross Quinlan. Improved estimates for the accuracy of small disjuncts. *Machine Learning*, 6(1):93–98, 1991.

[176] Vijay Raghavan, Peter Bollmann, and Gwang S Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[177] Sabry Razick, George Magklaras, and Ian M Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1):405, 2008.

[178] Guillaume Rigaut, Anna Shevchenko, Berthold Rutz, Matthias Wilm, Matthias Mann, and Bertrand Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17(10):1030–1032, 1999.

[179] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[180] Martin Rodbell. proteins in membrane transduction. *Nature*, 284:17, 1980.

[181] Burkhard Rost. Review: protein secondary structure prediction continues to rise. *Journal of structural biology*, 134(2):204–218, 2001.

[182] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[183] Luis Rueda, Sridip Banerjee, Md Mominul Aziz, and Mohammad Raza. Protein-protein interaction prediction using desolvation energies and interface properties. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 17–22. IEEE, 2010.

[184] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, 38(suppl 1):D497–D501, 2010.

[185] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[186] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451, 2004.

[187] Roger A Sayle and E James Milner-White. Rasmol: biomolecular graphics for all. *Trends in biochemical sciences*, 20(9):374–376, 1995.

[188] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[189] Andreas Schlicker and Mario Albrecht. Funsimmat update: new features for exploring functional similarity. *Nucleic acids research*, 38(suppl 1):D244–D248, 2010.

[190] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK, 1994.

[191] Rony Seger and Edwin G Krebs. The mapk signaling cascade. *The FASEB journal*, 9(9):726–735, 1995.

[192] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 40(1):185–197, 2010.

[193] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLoS computational biology*, 3(3):e42, 2007.

[194] Gajinder Pal Singh, Mythily Ganapathi, and Debasis Dash. Role of intrinsic disorder in transient interactions of hub proteins. *Proteins: Structure, Function, and Bioinformatics*, 66(4):761–765, 2007.

[195] Berend Snel, Gerrit Lehmann, Peer Bork, and Martijn A Huynen. String: a webserver to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18):3442–3444, 2000.

[196] Ryan W Solava, Ryan P Michaels, and T Milenković. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, 28(18):i480–i486, 2012.

[197] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

[198] Amelie Stein, Roland A Pache, Pau Bernadó, Miquel Pons, and Patrick Aloy. Dynamic interactions of proteins in complex networks: a more structured view. *Febs Journal*, 276(19):5390–5405, 2009.

[199] Simon Stobart and Mike Vassileiou. Introduction to php. In *PHP and MySQL Manual*, pages 7–11. Springer, 2004.

[200] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl 1):D561–D568, 2011.

[201] Charles Tanford and Jacqueline Reynolds. *Nature's robots: a history of proteins*. Oxford University Press, 2001.

[202] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, et al. The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):41, 2003.

[203] Richard D. Taylor, Philip J. Jewsbury, and Jonathan W. Essex. A review of protein-small molecule docking methods. *Journal of computer-aided molecular design*, 16(3):151–166, 2002.

[204] R Core Team et al. R: A language and environment for statistical computing. 2012.

[205] RDevelopment Core Team et al. R: A language and environment for statistical computing. *R foundation for Statistical Computing*, 2005.

[206] Garabet G Toby and Erica A Golemis. Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods*, 24(3):201–217, 2001.

[207] Peter Tompa. Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10):527–533, 2002.

[208] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, et al. Global mapping of the yeast genetic interaction network. *science*, 303(5659):808–813, 2004.

[209] Sabine Tornow and HW Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research*, 31(21):6283–6289, 2003.

[210] Olga G Troyanskaya, Mitchell E Garber, Patrick O Brown, David Botstein, and Russ B Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461, 2002.

[211] Athanasios Typas, Robert J Nichols, Deborah A Siegele, Michael Shales, Sean R Collins, Bentley Lim, Hannes Braberg, Natsuko Yamamoto, Rikiya Takeuchi, Barry L Wanner, et al. High-throughput, quantitative analyses of genetic interactions in e. coli. *Nature methods*, 5(9):781–787, 2008.

[212] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18(5):343–384, 2005.

[213] Zoltan Vajo, Lynn M King, Tanya Jonassen, Douglas J Wilkin, Nicola Ho, Arnold Munnich, Catherine F Clarke, and Clair A Francomano. Conservation of the caenorhabditis elegans timing gene clk-1 from yeast to human: a gene required for ubiquinone biosynthesis with potential implications for aging. *Mammalian genome*, 10(10):1000–1004, 1999.

[214] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[215] Konstantin Voevodski, Shang-Hua Teng, and Yu Xia. Spectral affinity in protein networks. *BMC systems biology*, 3(1):112, 2009.

[216] Christian Von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

[217] Christian Von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. String 7 recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, 35(suppl 1):D358–D362, 2007.

[218] Christian Von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(suppl 1):D433–D437, 2005.

[219] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[220] Albertha JM Walhout, Raffaella Sordella, Xiaowei Lu, James L Hartley, Gary F Temple, Michael A Brasch, Nicolas Thierry-Mieg, and Marc Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.

[221] Albertha JM Walhout and Marc Vidal. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, 24(3):297–306, 2001.

[222] Albertha JM Walhout and Marc Vidal. Protein interaction maps for model organisms. *Nature Reviews Molecular Cell Biology*, 2(1):55–63, 2001.

[223] James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[224] Theresa Weaver and Philip Hieter. Genome cross-referencing and xrefdb: implications for the identification and analysis of genes mutated in human disease. *Nat Genet*, 15:339–344, 1997.

[225] Chia-Lin Wei, Qiang Wu, Vinsensius B Vega, Kuo Ping Chiu, Patrick Ng, Tao Zhang, Atif Shahab, How Choong Yong, YuTao Fu, Zhiping Weng, et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124(1):207–219, 2006.

[226] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[227] Peter E Wright and H Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2):321–331, 1999.

[228] Gang Wu and Edward Y Chang. Kba: Kernel boundary alignment considering imbalanced data distribution. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):786–795, 2005.

[229] Yuliang Wu, Qiang Li, and Xing-Zhen Chen. Detecting protein–protein interactions by far western blotting. *Nature protocols*, 2(12):3278–3284, 2007.

[230] Hongbo Xie, Slobodan Vucetic, Lilia M Iakoucheva, Christopher J Oldfield, A Keith Dunker, Vladimir N Uversky, and Zoran Obradovic. Functional anthology of intrinsic disorder. 1. biological processes and functions of proteins with long disordered regions. *Journal of proteome research*, 6(5):1882–1898, 2007.

[231] Yuling Yan and Gerard Marriott. Analysis of protein interactions using fluorescence technologies. *Current opinion in chemical biology*, 7(5):635–640, 2003.

[232] Jihoon Yang and Vasant Honavar. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*, pages 117–136. Springer, 1998.

[233] Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006.

[234] Z Yang, WH Tang, Almas Shintemirov, and QH Wu. Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(6):597–610, 2009.

[235] KH Young. Yeast two-hybrid: so many interactions,(in) so little time... *Biology of reproduction*, 58(2):302–311, 1998.

[236] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

[237] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[238] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM, 2001.

[239] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[240] Xin Zhou and Zhen Su. Easygo: Gene ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC genomics*, 8(1):246, 2007.

[241] Hongbo Zhu, Francisco S Domingues, Ingolf Sommer, and Thomas Lengauer. Noxclass: prediction of protein-protein interaction types. *BMC bioinformatics*, 7(1):27, 2006.

Appendix A

Publications

In this chapter I present the major results of my work during my PhD preiod, which are four scientific publications in peer-reviewed journals. Originally a cumulative thesis was planed; however acceptance notification for the second first author publication came in to late, and thus I decided to compose a traditional PhD thesis.

A.1 Prediction of protein interaction types based on sequence and network features

Florian Goebels and Dmitrij Frishman
BioMed Central Systems Biology, 7, 1-18, 2013

Protein protein interactions are key players in many biological processes, which leads to the fact that there are detailed protein inteaction maps for several model organisms. However, due to the nature of high-throughput experimental methods, almost all of those interaction maps contain only information whether or not two proteins form a complex. This binary readout gives no clue as to how strongly those protomers interact with each other, or how long the interaction last. This Knowledge about the lifetime and binding affinity of non-covalent protein assemblies is crucial for understanding their role in cellular processes.

In this work we focesed on two protein interaction classes, based one their lifetime, and spatio-temporal distribution. In case of lifetime we distinguish between obligate and non-obligate interactions dependent on whether or not the protomers can exist independently. As for spatio-temporal distribution an protein interactions can be

either simultaneously possible (SP) or mutually exclusive (ME). Simultaneously possible interactions possesses a unique binding site for each interaction partner allowing each of it's parter to bind at the same time, while mutually exclusive interactions bind their partners at different times via the same interface. So far classifier for predicting protein interaction types exploited known differences in binding interfaces derived from known three-dimensional structures of protein complexes. Thus, those methods are only applicable for protein interaction where a 3D structure is available, which is only a neglactable fraction of the currently measured intractome.

Here we created the PiType protein interaction classification pipeline, which allows an accurate 3D structure indipendent classification of known protein protein interaction into simultaneously possible (SP) and mutually exclusive (ME) as well as into obligate and non-obligate. In contrast to privious methods our method relies on network and sequence based information for classifying protein intereacion types, and thus it can be used to classiffy PPIs in a large scale manner. In addition our classifier achieves at least 80% F-measure and AuROC, which is only marginaly worse than it structure based counterparts. Furthermore, we conducted an intesive analysis of non-structure based features for both obligate/non-obligate and SP/ME protein protein interactions. We revealed that proteins in non-obligate tend to have larger disordered regions, more short linear motifs, and share a lower functional similarity to their parter than obligate interactions. As for SP/ME interactions, we showed that they are characterized by significant differences in network topology.

The study design and data analysis was conducted by Dmitrij Frishman and me. I did the programming and preformed the research. The paper was drafted by myself and Dmitrij Frishman.

A.2 PiType 2.0: a Web server for classifying protein interactions

Florian Goebels and Dmitrij Frishman
BMC Boinformatics, submitted, 2014

In this paper we present an updated version of our PiType pipeline: PiType 2.0. We had two main goals we wanted to achieve in the updated version.

Firstly, we wanted to increase the total number of interactions, and species our classifier can process, due to numerous feedback from scientific community requesting different species than the three available species (Human, Yeast, *e. coli*) in the original PiType. This was achieved by integrating the STRING 9.0 database, which consist of a total of 336561678 interactions from more than 1100 species. However, due to the size of the STRING 9.0 network several changes were needed, which increased the speed of the pipeline by four fold.

Secondly, we wanted to increase accessibility of the PiType pipeline to enable convenient usage. This was achieved by implementing an easy to use web service for classifying protein protein interactions, which is publicly available under `http://webclu.bio.wzw.tum.de/PiType/index.php`. PiType 2.0 does not require any installation and ensures optimal runtime via using our in house computational grid with more than 200 nodes. The server itself need the following inputs: I) The user need to select which set of features he wants to use (Degree, EGDV, Disordered regions, ELMS, and/or Functional Similarity), II) the user needs to submit its interactions of interest, which can be uploaded as a text file or inserted in a text field, and III) the user needs to select which species network he wants to use. Optionally, the user can submit his email and will be informed when the calculations are done.

Me and Dmirtij Frishman developed and designed the new updates for PiType. I was responsible for the implementation. The paper was drafted by both authors.

A.3 Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis

Philipp Blohm, Goar Frishman, Pawel Smialowski, **Florian Goebels**, Benedikt Wachinger, Andreas Ruepp, and Dmitrij Frishman
Nucleic acids research, 42(D1), D396-D400, 2013

Concerning protein protein interaction prediction methods, a set of non interacting protein interactions (NIPs) is equally important then the set of positive protein protein interactions. There are several methods for creating a set of NIPs, however there is no know "gold standard" method for creating NIPs. Thus, in an effort to create a reliable set of NIPs, we created the Negatome 2 database, which is an expert generated set of NIPs. The database consists of two parts, one part of manually evaluated NIPs, and one part of NIPs which were inferred from structure based data taken from the PDB database. At first we used a text mining based method to screen all available literature for candidate NIPs, which in turn were manually validated by Goa Frishman. Thus, creating 2171 NIPs from the available literature, which I joined with the 4397 NIPs taken from the PDB and thus creating the total Negatome 2 data set of 6532 NIPs. We created the stringent Negatome 2 where we removed all NIPs which were reported to interact according the iRefIndex database, thus after the filtering step a total of 1991, 4161, and 6136 NIPs remained for the text mining, structure bassed, and total data set, respectively.

The work in this project was distributed so that Philipp Blohm created text mining based NIPs, which in turn were manually validated by Goa Frishman, Pawel Smialowski created the structure based NIPs, Florian Goebels filtered the new data and merger it with the old Negatome database, and the project was designed and supervised by Dmitrij Frishman. The following people drafted the manuscript: Dmitrij Frishman, Goa Frishman, Philipp Blohm, Pawel Smialowski, and Florian Goebels. Each author read and approved the paper draft.

A.4 Estimation of relative effectiveness of phylogenetic programs by machine learning.

Mikhail Krivozubov, **Florian Goebels**, Sergei Spirin
Journal of bioinformatics and computational biology, 12(2), 4.2014

This publication originated from my four month exchange stay with the Moscow State University, which was sponsored and supported by the RECESS graduate school. In this publication our main mission was to create a machine learning method which is capable of predicting the quality of phylogeny reconstruction basing on features, which can be calculated from the input alignment. In this work we focused on two common methods: Fitch-Margoliash (FM), and Unweighted Pair Group Method with Arithmetic Mean (UPGMA). For the machine learning prediction we used a random forest classifier, which we trained with alignments from an orthologous series (OS), for which the phylogeny reconstruction could be evaluated. In this publication we showed that the quality of phylogeny reconstruction can be predicted with more than 80% precision. In the next step we trained the classifier method to evaluate which phylogeny reconstruction method is better for a given alignment. However, with the previously used set of features we only correctly predicted 56% of the cases in which UPGMA was the superior method. This suggest that our method is marginally better than a random predictor, but if we take into account that UPGMA was better than FM in only 34% of the training cases, we can assume that our predictor is better than a random classifier.

In this work Sergei Spiring, and Mikhail Krivozubov created the training data and the test sets, also they created the methods and features used for classification. I contributed in training and evaluating the random forest classifiers, as well as measuring the importane of every used feature. The paper was drafted by every author.

List of Abbreviations

APR — PageRank affinity
auROC — area under receiver operating curve EGDV — edge graphlet degree vector
ELM — short linear eukaryotic motifs
GO — gene ontology
KNN — k nearest neighbors
MCM — mini-chromosome maintenance
ME — mutually exclusive
PR — precision-recall
ROC — receiver operating characteristic
SIN — structural interaction network
SP — simultaneously possible
SVM — support vector machine
TAP — tandem affinity purification
Y2H — yeast two hybrid

Acknowledgements

I would like to thank my family and all my friends for supporting me the last three years. Also I would like to thank all my colleges including the ones from Moscow for having interesting discussions as well as their help, and advice during my doctoral studies.

I would like to thank Prof. Dmitrij Frishman for giving me the opportunity to conduct my PhD thesis at the department of genome-oriented bioinformatics at the Technical University of Munich. Also I would like to thank Prof. Zimmer for his supervision of the thesis.

Special thanks go to Claudia Luksch who helped and encouraged me to do an exchange with the Lomonosov Moscow State University, and also giving good advice and counseling.

Special thanks go to Sergei Spirin for welcoming and collaborating with me during my stay at Lomonosov Moscow State University.

Last but not least I would like to thank Leonie Lorrie for all her help with administrative problems.

Curriculum Vitae

Curriculum Vitae — Florian Goebels

Persönliche Informationen

Adresse: Naumannstrasse 6 80997 München ⋄ Email: florian.goebels@googlemail.com ⋄ Geburtsdatum: 08.01.1986
Nationalität: Deutsch

Kernkompetenz:

- Statistischer Datenanalyse, Maschinellen Lernen und Datenbankauswertung.
- Analytisches Denken, sowie kreatives und systematisches Problemlösen.
- Gute Kommunikations- und Teamkompetenzen.

Ausbildung

7/2011-jetzt	**Doktor der Naturwissenschaften** Technische Universität München am Lehrstuhl für Genome Orientierte Bioinformatik Betreuer: Prof. Dr. Dmitrij Frishman Titel: Classification of protein-protein interaction using sequence and network based features
10/2005-9/2010	**Diplom Bioinformatik** Schwerpunkte: Graphentheorie und Maschinelles Lernen Ludwig-Maximilians-Universität und Technische Universität München Abschluss: Diplom Bioinformatik **Diplomarbeit** Betreuer: Prof. Dr. Ralf Zimmer Titel: Network orientated transfer of experimental data
9/1996-7/2005	Gymnasium Dr. Florian Überreiter, München Abschluss: Abitur

Arbeitserfahrung

9/2012-11/2012 und 4/2013-5/2013	**Internationaler Austausch mit Moscow State University** Betreuer: Dr. Sergei Spirin Während meines Besuches habe ich mit Dr. Spirin und seiner Forschungsgruppe zusammen gearbeitet. Wir haben eine Methode entwickelt, die Qualität von phylogenetischen Bäumen vorher sagt.
5/2012-8/2012 und 8/2011-4/2012	**Betreuung von Bachelor- und Masterarbeit** Ich habe sowohl einen Bachelor als auch einen Master Student bei der Fertigstellung seiner Abschlussarbeit betreut.

Stipendien

7/2011-7/2014	Regulation and Evolution of Cellular Systems (RECESS) Stipendium

Lehre

Herbst 2011/2012/2013	**Bioinformatik für Biowissenschaftler I**
Herbst 2011/2012/2013	**Master Praktikum: genome orientated bioinformatics**
Frühling 2012	**Bioinformatik für Biowissenschaftler II**

Außerschulische Aktivitäten

8/2012-9/2013	**Doktorandenvertreter** Offizieller Doktorandenvertreter an der Technischen Universität München.
4/2013-5/2013	**Mitglied des Munich Interact Symposium** Ich war verantwortlich für die Organisation und Umsetzung eines internationalem Doktoranden Symposium mit fast 400 Teilnehmer.

Seminare

7/2012	**Statistical Data Analysis II**
2/2013	**Scientific Paper Writing**
2/2012	**Time Management: Plan your Time Efficiently**
1/2012	**Statistical Data Analysis I**
10/2011-2/2012	**Advanced Writing Practice**
12/2011-1/2012	**Professional Leadership in Project management**

Programmierfähigkeiten

Betriebssysteme	Linux , Mac OS, sowie Microsoft Windows
Verteilte Systeme	Sun Grid Engine
Datenbanken	Design und MySQL
Programmier Sprachen	Java, Python, Perl, R, Shell scripting, HTML, PHP sowie Javascript

Sprachen

English	Verhandlungssicher
German	Muttersprache
Russisch	Grundkenntnisse

Referenzen

Dmitrij Frishman	Professor am Lehrstuhl für Genome Orientierte Bioinformatik (d.frishman@wzw.tum.de)
Ralf Zimmer	Dekan Ordinarius für angewandte Informatik mit Fokus auf Bioinformatik an der LMU München (Ralf.Zimmer@bio.ifi.lmu.de)
Sergei Spirin	Gruppenleiter am Belozersky Institute of Physico-Chemical Biology (sas@belozersky.msu.ru)

Präsentationen

Posterpräsentation:

2/2012	**Florian Goebels**, and Dimitrij Frishman, Classification of Protein Interactions, Munich Interact Symposium (Munich)
7/2013	**Florian Goebels**, and Dimitrij Frishman, Classifying Protein–Protein Interactions, ISMB/ECCB (Berlin)
7/2013	Mikhail Krivozubov, **Florian Goebels**, and Sergei Spirin, Estimation of relative effectiveness of phylogenetic programs by machine learning, MCCMB (Moscow)
7/2014	**Florian Goebels**, and Dimitrij Frishman, PiType 2.0: Protein interaction type prediction server, ISMB (Boston, MA)

Vorträge:

12/2013	**Florian Goebels**, and Dimitrij Frishman, Prediction of protein interaction types based on sequence and network features, GIW (Singapore)

Publikationen

1. **Florian Goebels**, and Dimitrij Frishman, Prediction of protein interaction types based on sequence and network features. BMC Systems Biology, 2013. 7(6): p. 1-18.
2. Philipp Blohm, Goar Frishman, Pawel Smialowski, **Florian Goebels**, Benedikt Wachinger, Andreas Ruepp and Dmitrij Frishman, "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis."Nucleic acids research 42.D1 (2014): D396-D400.
3. Mikhail Krivozubov, **Florian Goebels**, and Sergei Spirin, Estimation of relative effectiveness of phylogenetic programs by machine learning, JCB, accepted
4. **Florian Goebels** and Dmitrij Frishman, PiType 2.0: Protein interaction type prediction server, submitted
5. **Florian Goebels**, Robert Küffner, Quibin Luo, and Dmitrij Frishman, DIMA 4.0: Domain Interaction Map, manuscript in preparation