

Computational Methods for Integrating and Analyzing Human Systems Biology Data

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich–Technischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Hagen Blankenburg

Saarbrücken
Januar 2014

Tag des Kolloquiums:	28.5.2014
Dekan der Fakultät:	Prof. Dr. Markus Bläser
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Hans-Peter Lenhof
Berichterstatte:	Prof. Dr. Mario Albrecht
	Prof. Dr. Volkhard Helms
Akademischer Beisitzer:	Dr. Olga Kalinina

Abstract

The combination of heterogeneous biological datasets is a key requirement for modern molecular systems biology. Of particular importance for our understanding of complex biological systems like the human cell are data about the interactions of proteins with other molecules. In this thesis, we develop and apply methods to improve the availability and the quality of such interaction data. We also demonstrate how these data can be used in interdisciplinary studies to discover new biological results.

First, we develop technical systems for the instant integration of interaction data that are stored and maintained in separate online repositories. Second, we implement a computational framework for the application of multiple scoring algorithms to qualitatively assess different aspects of interaction data. Our methods are based on distributed client-server systems, ensuring that the services can be updated continuously. This promotes equal access to interaction data and allows researchers to expand the client-server systems with their own service.

Third, we focus our application studies on integrative network-based analyses of human host factors for viral infections. Our applications provide new biological insights into the life cycle of the hepatitis C virus and identify new potential candidates for antiviral drug therapy.

Kurzfassung

Die Kombination verschiedener biologischer Datensätze ist für die moderne molekulare Systembiologie unumgänglich. Eine besondere Bedeutung für unser Verständnis von komplexen biologischen Systemen wie der Zelle haben dabei Daten über die Wechselwirkungen von Proteinen mit anderen Molekülen. In dieser Arbeit entwickeln und verwenden wir Methoden zur Verbesserung der Verfügbarkeit und Bewertbarkeit von solchen Interaktionsdaten. Wir zeigen auch, wie diese Daten in interdisziplinären Studien genutzt werden können, um neue biologische Erkenntnisse zu gewinnen.

Zuerst entwickeln wir technische Systeme, um Interaktionsdaten von verschiedenen Quellen des Internets zusammenzuführen. Danach entwickeln wir ein computergestütztes System, welches die Anwendung verschiedener Algorithmen ermöglicht, um unterschiedliche Aspekte von Wechselwirkungen qualitativ zu bewerten. Unsere Methoden basieren auf verteilten Client-Server-Systemen, die sicherstellen, dass einzelne Dienste dauerhaft aktuell gehalten werden können. Zudem fördert dies einen gleichberechtigten Zugang zu Interaktionsdaten, und Wissenschaftler können die Systeme mit eigenen Diensten erweitern.

Unser Anwendungsschwerpunkt liegt auf der netzwerkbasierten Analyse humaner Wirtsfaktoren für virale Infektionen. Unsere Auswertungen tragen zu einem besseren Verständnis des Lebenszyklus des Hepatitis-C-Virus bei und zeigen Ansatzpunkte für die Entwicklung neuer antiviraler Medikamente auf.

Acknowledgements

The work presented in this thesis has been carried out in the research group *Molecular Networks in Medical Bioinformatics* in the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics*. I am much obliged to Thomas Lengauer for giving me the opportunity to work in this great and inspiring place.

I would like to express my particular gratitude to my advisor Mario Albrecht for providing invaluable support, motivation, and guidance throughout the years, for trusting me to pursue my own research ideas and offering precious feedback. I am also grateful to Volkhard Helms who kindly agreed to referee this thesis and to Hans-Peter Lenhof and Olga Kalinina for chairing and recording my colloquium, respectively.

This work would not have been possible without a large number of collaboration partners and co-authors. I am especially thankful to Andreas Prlić, Rob Finn, Andrew Jenkinson, and Bruno Aranda for helpful discussions on computational topics and to Marion Poenisch and Ilka Rebhan for patiently sharing their knowledge of virology.

Special thanks to all former MPII colleagues for creating such a lively and friendly work environment, in particular, Sven-Erik Schelhorn, Fidel Ramírez, Dorothea Emig, Andreas Schlicker, Nadezhda Doncheva, Nora Speicher, Adrian Alexa, Kasia Bozek, Konstantin and Laura Halachev(a), Jasmina Bogojeska, and Yassen Assenov. Unfortunately, I cannot mention all the people that made Saarbrücken a surprisingly nice place to work and live. A big thank-you to Joachim Büch, Georg Friedrich, and Ruth Schnepfen-Christmann for making sure the department and all its infrastructure run smoothly.

Finally, I would like to heartfully thank Eva and my family for all their understanding, love, and support.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Overview	2
1.3	Outline	3
2	Basics of molecular interactions	5
2.1	Molecular interaction types	5
2.2	Determination techniques	10
2.2.1	Experimental detection	10
2.2.2	Computational prediction	11
2.3	Interaction databases	12
2.3.1	Primary interaction databases	12
2.3.2	Aggregate databases	15
2.4	Interaction data quality	15
2.5	Interaction data formats and standards	16
2.5.1	HUPO-PSI-MI	16
2.5.2	PSI-MI 2.5 XML format	17
2.5.3	MITAB format	18
2.5.4	Binary representation of protein complexes	20
3	Data decentralization in interactomics	23
3.1	Introducing the Distributed Annotation System	23
3.1.1	DAS architecture	24
3.1.2	DAS protocol	27
3.2	DAS for Molecular Interactions	33
3.2.1	DAS protocol extension	34
3.2.2	DASMI server	40
3.2.3	DASMI client	42
3.3	PSI Common Query Interface	52
3.4	Conclusions	55
3.4.1	Summary	55
3.4.2	Outlook	57

4	Assessing molecular interaction data quality	59
4.1	Introduction	59
4.1.1	Interaction data quality issues	59
4.1.2	Methods for assessing molecular interactions	61
4.2	Decentralized scoring systems	66
4.2.1	Confidence scoring with DASMI	66
4.2.2	PSI Common Confidence Scoring System	69
4.3	Conclusions	76
4.3.1	Summary	76
4.3.2	Outlook	77
5	Integrative network analyses of viral host factor screens	79
5.1	Introduction	79
5.1.1	Hepatitis C virus	79
5.1.2	Host factor determination using RNA interference	82
5.2	Computational framework for analyzing viral RNAi screens	85
5.2.1	Datasets	85
5.2.2	Functional network analysis of viral host factors	87
5.3	Application studies	89
5.3.1	Human kinome screen for host factors required in HCV entry and replication	90
5.3.2	Human druggable-genome screen capturing all stages of the HCV life cycle	93
5.4	Conclusions	100
5.4.1	Summary	100
5.4.2	Outlook	101
6	Conclusions	103
6.1	Summarizing remarks	103
6.2	Perspectives	105
	Appendix	151
1	Technical documents	153
2	List of Abbreviations	175
3	List of own publications	179

List of Figures

2.1	Schematic drawing of different types and levels of molecular interactions	6
2.2	Network representation of the human interactome	9
2.3	Binary representations of protein complexes	20
3.1	Client-server architecture of the Distributed Annotation System	25
3.2	Client-server architecture of the Distributed Annotation System for Molecular Interactions	34
3.3	DASINT XML schema definition	38
3.4	DASMIweb graphical user interface	44
3.5	Querying and identifier mapping in DASMIweb	46
3.6	DASMIweb interaction view	47
3.7	Interaction details in DASMIweb	49
3.8	DASMIweb data source configuration	50
3.9	iPfam graphical domain interaction browser	52
4.1	Network representation of a high-confidence subset of the human in- teractome	63
4.2	Visualization of confidence scores in DASMIweb	68
4.3	PSISCORE architecture	70
4.4	PSISCOREweb graphical user interface.	75
5.1	Life cycle of the hepatitis C virus	82
5.2	Schematic experimental setup of RNA interference screens	84
5.3	Functional similarities of HCV host factors	89
5.4	Network representation of HCV host factors	92
5.5	Score distribution druggable primary screen	93
5.6	Functional annotation of hits in the druggable primary screen	95
5.7	Score distribution druggable validation screen	97
5.8	Integrated network visualization of host factors and functional anno- tation	98
5.9	Interaction network of HNRNPK	99

List of Tables

3.1	Versions of the Distributed Annotation System	24
3.2	DAS commands and XML response formats	29
3.3	DASMI sources providing domain and protein interactions	43
5.1	Published hepatitis C host factor screens	86
5.2	Host dependency factors validated in the kinome screen	91
5.3	Host factors confirmed in the HCV druggable screen	96
5.4	Selection of biological processes enriched among validated hits of druggable screen	97

Listings

3.1	XML representation of two DAS coordinate systems	27
3.2	Exemplary DASSEQUENCE XML document.	30
3.3	DASSEQUENCE schema as Document Type Definition (DTD)	31
3.4	DASSEQUENCE schema as XML Schema (XSD)	31
3.5	DASSEQUENCE schema as RELAX NG	32
3.6	DASGFF feature representation of a binary protein-protein interaction .	35
3.7	Syntax of the <code>interaction</code> DAS command	36
3.8	Exemplary DASINT XML document describing a binary protein-protein interaction with additional confidence scores	39

Für Mutsch – und die beiden, die nicht mehr hier sein können.

1 Introduction

This chapter introduces the topic of the thesis. First, some problems are detailed that motivated our research in biomedical data integration and its application to particular scientific problems in human systems biology. Then, the work that has been carried out during the course of the thesis will be presented, followed by an outline of the thesis structure.

1.1 Motivation

Over the last years, molecular biology has been undergoing a fundamental change, largely driven by the development or improvement of experimental techniques that allow studying diseases and associated molecular processes in ways that have not been possible before. For example, RNA interference screening, a technology that has only been established about a decade ago, now is capable of selectively investigating the effect of every single gene in our genome on certain disease-associated pathways or infections¹. This opens up exciting new avenues for basic research and clinical diagnostics, which were wholly unthinkable a few years ago.

The technological breakthroughs also impacted the general way research is carried out. The traditional reductionist and hypothesis-driven approach splits problems into sub-problems until they can be experimentally tested. Now, experiments may be performed in a less biased fashion, via *ome*-wide high-throughput techniques that are not limited to preselected study objects. Instead of investigating whether two particular proteins physically bind to another, researchers can test thousands of proteins or even complete proteomes, that is, all proteins of an organism, against each other in a single large-scale experiment.

This new way of approaching research questions in molecular biology is also reflected in the paradigm of systems biology, an interdisciplinary research area that tries to analyze and model cells and other complex biological units as highly interconnected components that need to be investigated as a whole. Different *omics* disciplines evolved to this end, for example, *proteomics*, which investigates the proteome, *lipidomics*, which researches connections between all cellular lipids, or *metabolomics*, which analyzes chemical processes and the molecules involved.

Vast amounts of heterogeneous data are produced by the multitude of high-throughput techniques in the different *omics* disciplines of systems biology. Analyz-

ing these data may easily be beyond the capabilities of the individual experimenters, requiring expertise in areas like statistics and information processing. Bioinformatics is becoming increasingly relevant because it provides the link between biological knowledge and means to analyze large quantities of data systematically using computers and sophisticated algorithms.

Of central importance to this thesis is the area of *interactomics*, which investigates the interactions of proteins and other molecules. Proteins, the key players involved in virtually every cellular process, perform their functions through interactions with other molecules. A comprehensive understanding of complex cellular systems is thus not conceivable without knowing the interactome, the set of all molecular interactions in a cell⁷⁸. In addition, protein interactions emerge as potential targets for novel therapeutic approaches to combat disease^{328,429}.

Several issues are currently associated with protein interaction data, limiting their usefulness for systematic studies³⁹². Two particular problems are general data availability, as interaction data are not stored centrally, but spread over multiple online repositories, and data quality, as different determination techniques produce results that are difficult to assess and compare. In this thesis, I will present methodological improvements and new systems, which may help to ensure that interactome data are available to interested researchers in a convenient, high-quality, and timely manner.

Interactomics is not only concerned with measuring and modeling interaction data. Molecular interactions and the networks and pathways they form, are also an important resource for the applied, integrative bioinformatics analysis of different experimental datasets. For example, knowledge about the interaction partners of a protein may be used to infer information about its biological function. Incorporating additional, complementary datasets like expression profiles or functional annotations into such network-based analysis approaches, may allow detecting patterns in the experimental data that could not be uncovered by individual investigations.

1.2 Overview

The work performed while preparing this thesis can be divided into two parts. First, the development of methods and software to improve the availability and data quality of protein interactions. Second, the combination of molecular interactions and other types of biomedical data in application studies to answer particular research questions. Importantly, by closely working together with external research groups, expertise in biological domains is brought together with our bioinformatics knowledge.

The methodological developments in the first part consist of distributed systems that we implemented in collaboration with biomedical data providers. After showcasing the potential of data decentralization with our Distributed Annotation System for Molecular Interactions^{40,40,181}, we (co-)developed two follow-up projects in the context of the Proteomics Standards Initiative of the Human Proteome Organiza-

tion, building on experiences gained with the initial prototype¹⁴.

The application studies, which form the second part of the thesis, apply different techniques for the integration, analysis, and prioritization of heterogeneous biomedical data types in order to answer questions of direct biological or medical relevance. A long-lasting and fruitful collaboration was established with the Department of Molecular Virology at the Heidelberg University, which we supported with the analysis of several large-scale RNA interference screens to determine human host factors required for viral infections^{105,321}.

The work presented in this thesis resulted in six co-authored publications in peer-reviewed journals or conference proceedings and three book chapters. For the sake of brevity and clarity, the applied bioinformatics work that led to two additional co-authored publications^{127,423} is not described here in more detail.

1.3 Outline

The remainder of this thesis is split into five chapters followed by a bibliography and an appendix.

Chapter 2 introduces the basic types of interactions between proteins and other molecules, an area that is important for all other chapters. First, experimental and computational techniques for determining those interactions are presented. This is followed by descriptions of resources for storing and retrieving interaction data and the main data formats and standards.

Chapter 3 presents computational methods to improve the general availability of molecular interaction data. After introducing the Distributed Annotation System (DAS), our extension named DAS for Molecular Interactions (DASMI) is presented in detail. Today, DASMI is largely superseded by a successor system named Proteomics Standards Initiative Common Query Interface, which is briefly described in the remainder of the chapter.

Chapter 4 focuses on the evaluation of interaction data quality. After detailing the existing quality problems, an overview of various approaches to assess interaction data quality is provided. We then show how DASMI can be used for the integration and visualization of different scoring approaches. Based on DASMI, we develop the Proteomics Standards Initiative Common Confidence Scoring System that is described thereafter.

Chapter 5 contains applications of different data integration techniques to the analysis and prioritization of functional genomic screens determining human host factors for viral infections. After providing an introduction to the hepatitis C virus (HCV) and systems for studying host factors, I will describe the computational platform that we have developed for the analysis of such screens. Then, a network-based functional analysis of HCV host factors is presented and the application of our analysis platform to two RNA interference studies for determining HCV host factors is presented.

[Chapter 6](#) summarizes the thesis and provides an outlook on future advancements.

The [Appendix 6.2](#) or [Appendix 6.2](#) contains the technical documentations of the distributed systems we developed and a list of own publications.

2 Basics of molecular interactions

This chapter provides an introduction into the basic concepts and types of molecular and protein interactions, as this biological data type is relevant for all chapters of the thesis. The different molecular interaction types and the required terminology is introduced in [Section 2.1](#). In [Section 2.2](#), I will present several techniques for experimental interaction detection or for their computational prediction. The main resources for storing and retrieving interaction data are presented in [Section 2.3](#). Only a brief overview of interaction data quality is provided in [Section 2.4](#), as [Chapter 4](#) is dedicated to this topic. A description of the relevant standards, data formats, and guidelines in the field of molecular interactions is presented in [Section 2.5](#).

2.1 Molecular interaction types

Proteins are essential building blocks of all organisms that have manifold roles as molecular machines of our cells. Different groups of proteins are fundamental to virtually every aspect of cellular functioning, including membrane and signaling proteins that form cellular signaling pathways, structural proteins that stabilize cells or cellular compartments, enzymes that catalyze chemical reactions, transport proteins that deliver molecules, or immunoproteins that mediated the human immune response²⁹². Most of this functionality is not accomplished by individual proteins but via an interplay with other proteins or molecules like lipids, nucleic acids, or metabolites^{101,123}. An illustrative example is the adenosine triphosphate (ATP) driven interaction between the proteins actin and myosin, that ultimately leads to the contraction of muscle fibers and thus to our ability to move¹⁶⁸.

Proteins The interaction between proteins or other molecules is referred to as protein or molecular interaction. There is no strict definition for these terms and they are sometimes used synonymously. Throughout this thesis, I will use the term *protein interaction* for relationships that only involve proteins, while the term *molecular interaction* will be used as a more general, higher-level term that denotes interactions involving other molecules like nucleic acids, hormones, or drugs. Using this definition, protein interactions are a subset of molecular interactions.

Depending on the type and the number of interacting molecules, which in the following are also named *interactors*, protein interactions may be grouped into different

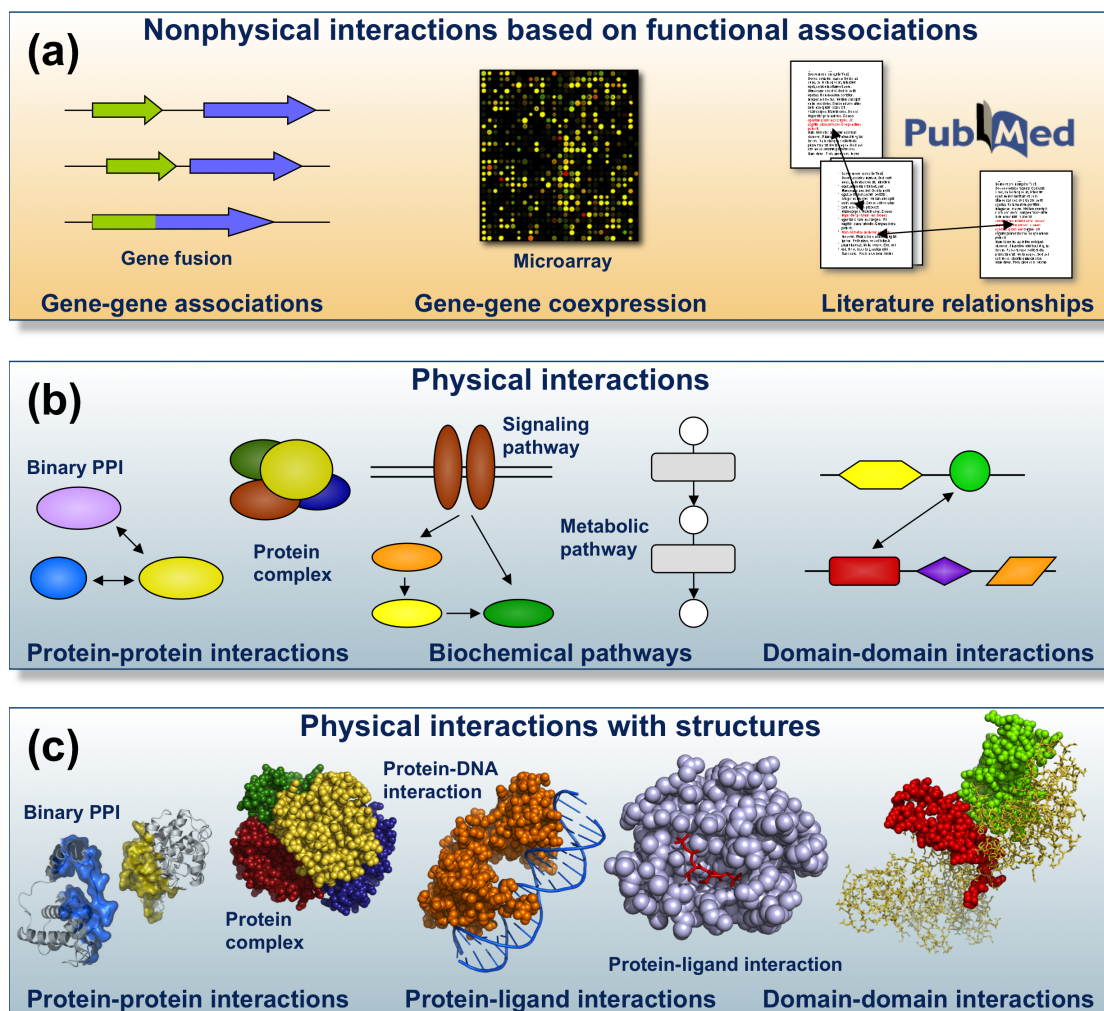


Figure 2.1: Schematic drawing of different types and levels of molecular interactions⁴⁰. In general, physical and non-physical interactions can be distinguished. (a) Non-physical interactions are also known as associations and describe a functional relationship between two entities, for example, the co-mentioning of two genes in a sentence of a scientific publication or their co-expression in a certain cellular condition. (b) Physical interactions imply an actual physical contact between the interacting entities (interactors). (c) By raising the level of detail, the interacting atoms or residues of the binding sites can be identified.

categories (see [Figure 2.1](#)). If up to two proteins are interacting, this is termed a *protein-protein interaction* (PPI) or *dimer*. The term *binary* may be added to emphasize the fact that exactly two proteins are involved. An assembly of more than two proteins is named *protein complex*, *oligomer*, or *n-ary* interaction. Depending on the identity of the interacting proteins, dimers and oligomers can be further divided into *homodimers* and *homooligomers* if identical proteins are interacting, and *heterodimers* and *heterooligomers* if different proteins are involved, respectively²⁶⁶. Throughout this thesis, I will use the term *protein interaction* to refer to both, binary PPIs (homo- and heterodimers) as well as protein complexes (homo- and heterooligomers). The similarly looking term *protein-protein interaction*, in contrast, will only be used when referring to binary interactions involving exactly two interactors, thus excluding protein complexes.

Domains and motifs Most proteins are modular and are composed of subunits like protein domains or short linear motifs^{258,292}. A *protein domain* is commonly defined as a unit comprising at least 30 amino acids that can fold stable and independent of the rest of the protein. Other definitions also include the functional and evolutionary role of protein domains^{44,253,292}. *Short linear motifs* (SLMs) are smaller units, usually consisting of only three to ten amino acids, that do not form independent structures and are usually found in disordered, unstructured regions of a protein^{258,263}. By increasing the level of analysis detail, protein interactions can often be traced to underlying interactions between protein domains (*domain-domain interaction* (DDI)^{116,287,307,358}) or to short linear motifs that bind to the domain or other parts of a protein (domain-motif interaction^{2,258,260}). With respect to the interaction lifetime, one can distinguish transient interactions and stable or permanent interactions. For example, interactions mediated by DDIs are more stable compared to transient domain-motif interactions²¹.

Ligands Non-protein interactors will in the following be named *ligands*, the interaction between a protein and such a molecule consequently *protein-ligand interaction* (see [Figure 2.1](#)). Ligands may refer to molecules of different types, ranging from large macromolecules such as nucleic acids (*protein-DNA interaction*⁸⁹ and *protein-RNA interaction*²⁰¹) to small compounds like drugs or hormones⁴³¹. Examples for such protein-ligand interactions are the muscular transport protein myoglobin, which reversibly binds an oxygen molecule¹⁸⁹, or the interactions of transcription factors, proteins that bind to specific DNA sequences and thereby regulate genomic information flow¹¹². The region within the protein that is binding to the ligand is named ligand-binding site²⁴⁰. In enzymes, these regions are termed active sites and the ligands that bind to these sites are called substrates²⁹².

Associations All the molecular interaction types described thus far are examples of *physical* or *direct interactions* that imply an actual contact between the interacting

entities, for example, the binding of protein interfaces. The molecular basis for these interactions are the formation and dissociation of non-covalent atomic interactions like hydrogen bonds, van der Waals, or electrostatic interactions²⁹². In contrast, *non-physical interactions* or *associations* describe functional relationships between molecules but do not imply a physical contact. Two proteins may be associated if they are mentioned in the same sentence of a scientific publication or if they are co-expressed in the same cellular condition or compartment²⁴³. Genetic interactions, such as synthetic lethality, where mutations in two genes produce a combined effect that is unexpected considering the individual mutational effects, are another example for such non-physical associations²³². It is obvious that physical interactions and functional associations are fundamentally different and should not be unintentionally mixed in analyses.

Interactome and interactomics In our current age of ”-ome” and ”-omics” terms²¹², it is not surprising that there is a catchy name for the complete set of all molecular interactions that may occur within a cell: the *interactome*³³⁵. Consequently, the research area that deals with molecular interactions is named *interactomics*¹⁹⁶. The determination of the complete human interactome has always been an important research aim in systems biology^{78,391}. Estimates about the size of the human interactome are diverse, ranging from 130 000 to 650 000 PPIs^{125,148,285,355,363,388,396}. As of December 2013, the integrative interaction database mentha⁵⁸ provides about 140 000 binary human PPIs (see [Figure 2.2](#)). However, a significant number of those interactions are likely to be false positives, that is, interactions that are wrongfully reported by a single experiment or publication and cannot be reproduced. In fact, [Venkatesan et al.](#) estimated in 2009 that only about 9 percent of the complete human interactome was known at the time³⁸⁸.

The singular term interactome and ”hairball” representations like in [Figure 2.2](#), may lead to the wrong impression that there is only one static set of interactions. However, the interactome is highly dynamic, both in terms of time and space³⁰⁵. Certain interactions may or may not occur under particular circumstances, for instance, when the cell is under stress. As proteins elicit their functions through interactions with other proteins¹⁰¹, it is obvious that interactions do not constantly or randomly occur, but only when a particular cellular function is mediated. In addition, not all interactions that could in theory take place, for instance, because two proteins have complementary, energetically favorable binding sites, can also take place in the living cell. In order to interact with each another, proteins need to be expressed at the same time and in the same cellular components, cells, tissues, or organs²²⁶. While the dynamic nature of the interactome is well accepted, the fact is still not adequately represented by predominant experimental detection techniques, databases, or analysis software^{305,345}.

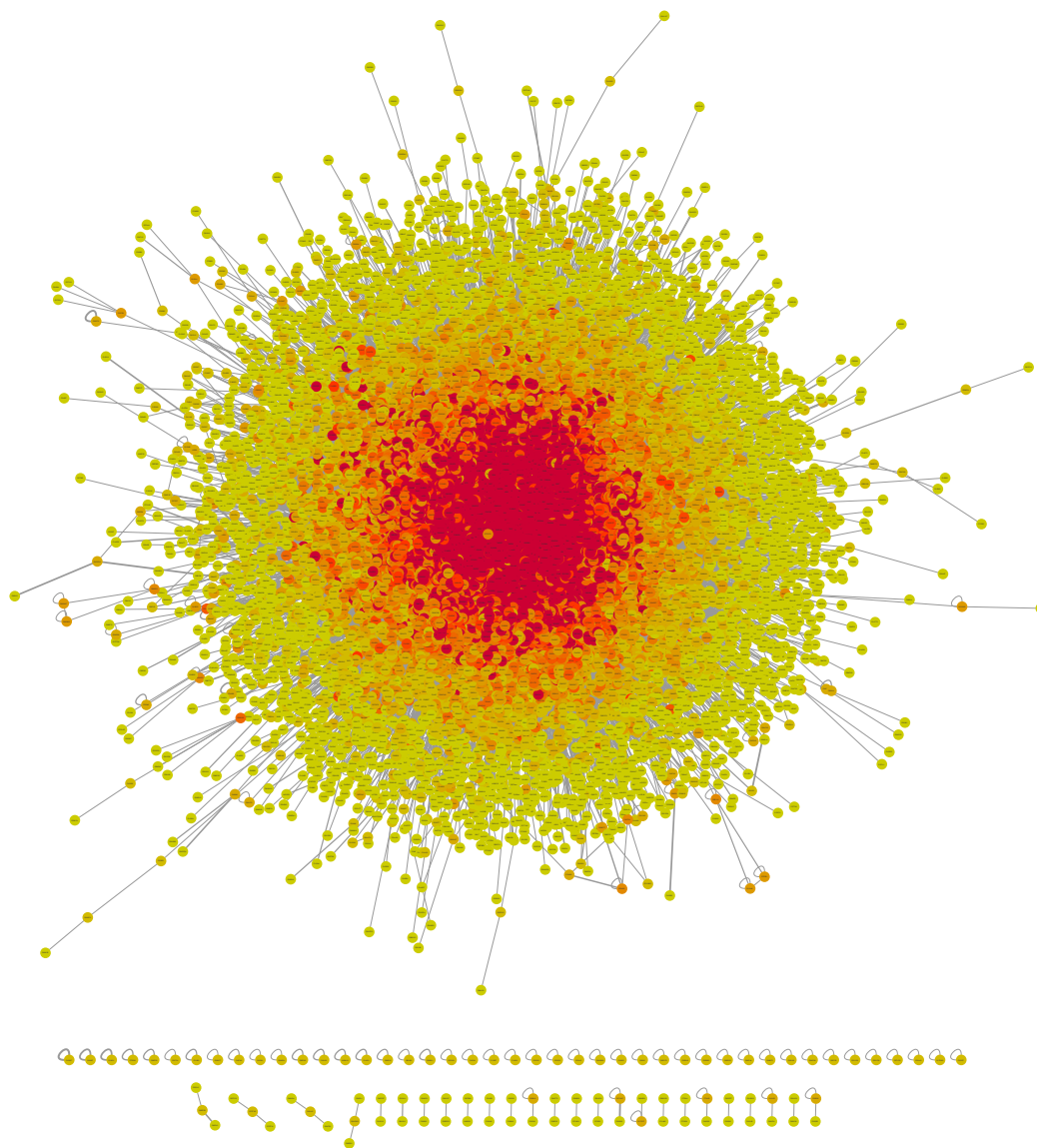


Figure 2.2: Network representation of the currently known human interactome. Nodes indicate proteins, edges depict binary PPIs. Protein complex data are not included. The network, retrieved from mentha⁵⁸ on 11 December 2013, consists of 14 650 proteins and 143 947 interactions. Visualization was done in Cytoscape 3.0.2 using a force-directed layout and a color gradient which places nodes with a higher degree (many interactions, colored in red) towards the center of the "hairball", whereas nodes with fewer interactions (green) are located towards the borders. The vast majority of nodes are in a large connected component.

2.2 Determination techniques

This section provides a brief description of the most important methods for determining potential molecular interactions. In general, the approaches for interactome mapping can be classified into experimental (*in vivo* and *in vitro*) and computational (*in silico*) techniques.

2.2.1 Experimental detection

Interactions between proteins have been analyzed experimentally even before the concepts of proteins, let alone their interactions, were clearly established⁴⁸. The traditional way, in line with the reductionist approach of classical biology³¹⁸, was to study one molecule at a time. Today, a great number of experimental approaches exist for determining protein and molecular interactions on different scales^{45,348}. The separation into small-scale or low-throughput and large-scale or high-throughput techniques is not always feasible, as some technologies may be used for a detailed characterization of individual protein interactions, but also for high-throughput studies⁴¹⁴. In addition, technological progress can scale up the throughput of a method significantly, moving it from low- to high-throughput¹¹⁴. Modern high-throughput techniques allow for screening complete proteomes of organisms in parallel³⁹². In the following, I will describe the two most widely used technologies for experimental interaction detection in a high-throughput fashion.

Yeast-two hybrid (Y2H) Yeast-two hybrid is a genetic system that has been developed more than 20 years ago for the detection of binary PPIs. Y2H takes advantage of some properties of the GAL4 system in yeast¹¹⁵. GAL4 is a transcription factor that is composed of two domains, a binding domain (BD) that binds to an enhancer named upstream activation sequence (UAS), and an activation domain (AD) containing acidic regions that are necessary to activate transcription of a downstream reporter gene¹¹⁵. By splitting these domains and fusing them to two different proteins (the protein fused to the binding domain is called *prey*, the protein fused to the activation domain *bait*), a physical protein-protein interaction can be detected, as transcription of the reporter gene will only take place if the two domains are brought into proximity. The individual domains will not activate transcription. Y2H started as a system for studying single PPIs¹¹⁴, but through continuous developments has evolved to a versatile platform that has successfully been used in a number of high-throughput studies in various organisms^{120,136,171,219,308,327,360,382}.

Affinity purification (AP) Affinity purification is a technique that uses specific tags to detect and "pull out" assemblies of proteins or other molecules. In the tandem affinity purification (TAP) protocol, TAP tags are fused to a target or *bait* protein. The tagged bait is introduced into a host cell where the protein may interact with

other *prey* proteins under physiological conditions³²⁴ (see [Figure 2.3](#)). Through a two-step purification process, the tagged protein and all molecules attached to it are then extracted, for example, from a lysed cell. The individual proteins that have been pulled out are then identified via mass spectrometry (MS). TAP-MS has been developed as a high-throughput technology, which has since been used in a number of large-scale studies^{56,129,130,204}. While the composition of protein complexes can be determined by TAP-MS experiments, their exact structures remains unclear and computational methods are needed to infer pairwise physical protein interactions within these complexes³³⁹. In order to represent n-ary protein complexes as binary PPIs, a number of complex expansion models have been proposed (see [Section 2.5.4](#)).

While the majority of large-scale interaction maps is determined using Y2H or TAP-MS, a number of additional technologies with lower throughput are commonly used for validation purposes³⁸⁸, including luminescence-based mammalian interactome mapping (LUMIER²⁵), mammalian protein-protein interaction trap (MAP-PIT¹¹¹) or fluorescence or Förster resonance energy transfer (FRET⁴¹⁴). It is important to know that each of the experimental techniques by design has certain strengths and weaknesses. For example, Y2H is capable of detecting interactions with very low affinities, while AP-MS requires a stronger binding between the proteins, as weaker links might be removed in the purifications process²⁰⁰. Generally, false positive and false negative interactions will occur with every technology^{49,230}. Wherever feasible, interactions are ideally experimentally determined using a variety of different approaches³⁸⁸.

2.2.2 Computational prediction

The experimental detection of molecular interactions is time-, work-, and cost-intensive, despite the establishment of high-throughput techniques like Y2H and TAP-MS³⁸⁸. As large parts of the human interactome still remain to be determined^{148,363}, a variety of computational methods have been proposed in order to predict potential interactions in the uncharted regions²⁰⁰. These *in silico* approaches will not replace experimental detection, which remains the only way of detecting true positive interactions. However, they might be used to narrow down the search space, complement experimental results, or even spot potential errors in experimental data⁴⁰³.

A popular *in silico* approach is to predict interactions based on different types of homology. In *interolog* mapping, an interaction is inferred between species via sequence homology, based on the assumption that if an interaction between two proteins A and B has been found in one species, most commonly a well-studied model organism like yeast or fly, and homologous proteins A* and B* exist in another species, for example, human, an interaction between A* and B* may be inferred^{51,128,237,419}. Homology is also used on a structural level for predicting interactions between proteins that have binding interfaces similar to those found in known interactions^{7,194,211,403,427}. Additional methodologies employ genomic or se-

quence features^{62,107,179,234}, orthology^{166,172}, phylogenetic profiles²⁸⁹ or combinations thereof^{31,109,386,426}. Computational text mining is also considered as a prediction technique^{197,200}. A comprehensive description of the different prediction algorithms is provided in numerous reviews^{45,225,294,349,384}.

2.3 Interaction databases

Interaction databases are resources that try to collect all the interaction data that are or have been produced in research laboratories around the world and make them available to the research community. The first dedicated public databases for human interaction data, among them the Database of Interacting Proteins (DIP)^{234,411} and the Biomolecular Interaction Network Database (BIND)^{17,19}, appeared at the end of the 1990s, in a time where the first large-scale protein interaction screens have been published⁴¹⁰. With the ever-growing attention in the research community, the number of available resources has since kept increasing. As of 2013, the pathway and interaction repository Pathguide¹⁸ lists well over 100 resources^a that, at least partly, provide interaction data. In practical terms, this number will be slightly smaller, as not all of these resources are still actively maintained and some are highly specialized, including those that list "proteins that interact with GroEL and factors that affect their release" (GroEL PPI, no longer available) or the "database of human GPCRs, G-proteins, Effectors and their interactions" (Human-gpDB³³⁶).

2.3.1 Primary interaction databases

BIND⁵ and DIP³³³ are examples of primary interaction databases, other popular examples are BioGRID⁶⁴, the Human Protein Reference Database (HPRD¹³⁸), IntAct¹⁹⁰, or the Molecular Interaction Database (MINT²²⁰). The term "primary" indicates that these are the first line between the raw interaction data and the user, in contrast to aggregate resources that pool together data from other databases (see Section 2.3.2). Many of the primary interaction databases have originally not been established with the aim of becoming general repositories, but have resulted from small, focused studies²⁷⁰. As a consequence, several of the primary resources listed in Pathguide, like the aforementioned GroEL PPI, BIND, or the microbial protein interaction database (MPIDB¹⁴⁰), are no longer actively maintained^{170,270}. A number of primary databases, however, have existed for more than a decade and have established themselves as important pillars for interactomics, among them DIP, IntAct and MINT, which have recently joined forces in the MIntAct project²⁷⁰.

Direct data submission Primary interaction databases generally have two ways to acquire new data. First, they receive interaction data as direct submissions from

^awww.pathguide.org/

experimentalists^{274,279}. Inspired by the microarray, sequencing, and protein structure fields, where it was agreed by major journals that only such manuscripts are published, where the referenced entities have been stored in a publicly accessible database before publication, a similar system has recently been established in the field of interactomics⁹⁸. Currently, however, not many journals are participating in this initiative. In addition, guidelines named *minimum information required for reporting a molecular interaction experiment* (MIMIx) have been developed to help experimentalists in providing the primary databases with all the information necessary for describing a molecular interaction in detail^{279,281}. Direct data submission is more common for dedicated interactome studies, for instance, large-scale Y2H or TAP-MS screens that yield hundreds of interactions and is promoted by journals that commonly receive such datasets⁶¹.

Text mining and data curation Very often, however, interaction studies between proteins are only performed as means of testing a certain biological hypothesis. For example, in order to characterize the molecular role of the human protein kinase phosphatidylinositol-4 kinase III alpha (PI4KIII α) in hepatitis C virus (HCV) infections (see Section 5.3.1), Reiss et al.³²¹ have used a small-scale co-immunoprecipitation experiment to test for interactions between PI4KIII α and all HCV proteins. The two interactions they found between PI4KIII α and the viral proteins NSA5A and NSA5B were consequently described in the publication and illustrated in the corresponding figures³²¹. However, they were not directly submitted to any of the primary interaction databases.

Therefore, another way of incorporating interaction data is by mining the scientific literature, an invaluable resource where interactions have been described as parts of experiments for decades, if not centuries⁴⁸. However, the incredibly large body of text makes this an enormous challenge. The MEDLINE/PubMed database^b, for example, currently contains more than 20 million entries spanning more than two centuries.

Computational text mining In general, current text-mining approaches can be divided into manual and computational techniques. Computational text-mining tries to automatically detect biological entities and relationships between them, for instance, using natural language processing³¹⁶. This is a very powerful technique that enables the analysis of huge bodies of text, like the complete PubMed database.

Without going into the details, the challenges of text mining can be nicely illustrated by the aforementioned interaction between PI4KIII α and the HCV proteins, which was described in the sentence: "We found that NS5A and NS5B, but not the NS3/4A protease/helicase or NS4B, interacted with the kinase."³²¹ For a human, this sentence is not overly complicated, but computational algorithms are faced with

^b<http://www.ncbi.nlm.nih.gov/pubmed>

several potential pitfalls. First, NS3 and the other viral proteins need to be identified as HCV proteins, not as proteins of the same name from Dengue or other viruses. Second, PI4KIII α is not clearly described by a name, database identifier, or accession number, but only described as "the kinase". Third, the two interactions that have explicitly been identified as not taking place might be mistaken for true interactions if the "not" keyword is not appropriately considered.

While the number of false positive and false negative interactions may be significant, several approaches utilize computational text mining for detecting molecular interactions in the scientific literature^{155,177,197,310}.

Manual text mining Fully automated computational text-mining is currently not considered a technique for identifying experimentally determined interactions but is seen as a form of computational prediction²⁰⁰. Primary interaction databases, in contrast, use a manual text-mining approach¹³. Curators are employed to read scientific articles and extract detailed interaction records. During peak times, for example, the Biomolecular Interaction Network Database (BIND) employed 40 curators¹⁷⁰. Compared to automatic text mining, manual curation has a significantly lower throughput but a much higher accuracy³³². In addition, the interaction records can be far more detailed, as the curator comprehends the complete article and can thus detect parts that describe experimental conditions, cell lines, study systems, or protein modification. In the PI4KIII α example, another sentence described the experimental conditions: "Individual HCV proteins were therefore coexpressed with HA-tagged PI4KIII α in Huh7-Lunet/T7 cells, and coimmunoprecipitation experiments were performed with monospecific antisera"³²¹.

IMEx consortium Particularly noteworthy in this regard is the IMEx consortium^c, which has formed to coordinate and improve the cooperation between primary interaction databases²⁷⁸. Before forming the IMEx consortium, the databases did not coordinate their text-mining endeavors, with the result that identical publications were potentially curated multiple times in different databases. As each of the databases used other curation standards, the same interactions could even be reported in different ways²⁷⁰.

In contrast, the participating IMEx databases have now agreed on common standards and procedures for curating interaction data. As a result, an interaction record in DIP should now contain the same level of information as an interaction record in MINT, as both follow IMEx curation guidelines. In addition, IMEx databases have assigned responsibilities for certain journals to each contributing database, preventing re-curation of the same article by different databases. In order to harmonize the data that are contained in the databases, the consortium has furthermore agreed on exchanging data on a regular basis²⁷⁸. Recently, several IMEx databases including IntAct and MINT have even gone further and decided on a common computational

^c<http://www.imexconsortium.org/>

platform that is directly filled by curators from the different projects, reducing potential redundancies to a minimum²⁷⁰. As of December 2013, eleven databases are participating in the IMEx consortium.

Despite the joint efforts, coverage will still remain an issue for the majority of organisms. While BioGRID claims to have reached complete coverage of the interaction literature for yeast⁶⁴, the protein interactions of PI4KIII α described by Reiss *et al.*³²¹, which I have used in several examples, are not yet available in any interaction database.

2.3.2 Aggregate databases

Primary interaction databases are essential components of interactomics, as they provide the most recent interaction data with the highest level of detail^{92,199}. However, manual curation of the scientific literature is time-intensive and the majority of these resources restrict their focus to experimentally determined interactions²⁷⁷. Despite harmonization approaches like the aforementioned IMEx consortium, the coverage of a single interaction database is still remarkably low and each of the databases contains unique information that is not found in the other resources^{236,313,381}.

Aggregate databases try to improve this situation by merging interaction data from various primary resources into a single repository. A significant number of the resources listed in Pathguide are aggregate database. Examples for protein interactions are the Agile Protein Interaction DataAnalyzer (APID²⁹⁸), HINT⁸², HitPredict²⁸⁴, iRefIndex³¹⁵, mentha⁵⁸, Michigan Molecular Interactions (MiMI¹⁸⁰), the Protein Interaction Network Analysis (PINA) platform⁷⁷, the Search Tool for the Retrieval of Interacting Genes (STRING³⁶⁹), or the Unified Human Interactome (UniHI⁶⁷). Aggregate databases that focus on domain-domain interactions include the Domain Interaction Map (DIMA²²⁷) or DOMINE³⁰⁷. Several resources like iRefIndex, STRING, or DOMINE combine computational predictions with experimentally determined interactions, while resources like UniHI or MiMI focus on experimental data. Compared to the source interaction entries, the records stored in aggregate databases usually do not maintain the same level of detail.

An important characteristic of almost all aggregate databases is that they perform a static unification of the interaction data into central repository³⁴⁶. That is, these databases represent a snapshot of the interactome at a certain point of time. Further efforts and short update cycles are required to assure that the data are kept in sync with the primary databases, as these are usually constantly evolving⁵⁸.

2.4 Interaction data quality

Data quality has been recognized as an important issue in the field of interactomics, ever since the first large-scale studies have been published. In particular, the low overlap between studies that had been performed independently in the same organ-

ism and with the same experimental technique raised questions about interaction data reliability^{86,171,382,396}. In the years since, a lively debate has formed in the scientific community about the reliability of certain experimental procedures^{230,388} or literature curation^{80,332}, which has even led to general prejudices and premature conclusions¹³³. Due to the great importance of data quality, this subject is described in more detail in [Chapter 4](#).

2.5 Interaction data formats and standards

The scientific literature is the traditional way of reporting the results of interaction experiments⁴⁸. As I have shown in the [Reiss et al.](#)³²¹ example (see [Section 2.3.1](#)), information is either textually described or placed in tables or figures. With the emergence of high-throughput screening, which can result in hundreds or thousands of reported interactions, this information was increasingly moved to dedicated text files or tables. In addition, the information is collected and made available by interaction databases. Until a few years ago, each of the files, tables, or databases used different specifics to represent the data¹⁵³. As a consequence, researchers had to constantly adapt to new presentation standards and develop dedicated parsers when using interaction data from various sources in their work.

2.5.1 HUPO-PSI-MI

In 2001, the Human Proteome Organization (HUPO) was established as an international collaboration to facilitate proteomics research¹⁴⁵. Within this organization, the Proteomics Standards Initiative (PSI) was formed with the aim of developing and promoting community standards for the representation and exchange of proteomics data^{152,275}. One of the standards that was developed under its guidance is the Molecular Interactions (MI) standard, which was published by a consortium of the major interaction data providers and primary databases in 2004¹⁵³. Version 1.0 of the PSI-MI standard consists of an Extensible Markup Language (XML) format definition (see [Section 2.5.2](#)) and a controlled vocabulary. The initial adaptation was diverse, while databases that were involved in designing the standard, such as IntAct, DIP, or MINT, offered data in this format, smaller databases did not support the format for some time³⁶². In addition, the standard was often interpreted differently by the various databases, with the result that PSI-MI files from different databases were not fully compatible^{37,315}. One factor that contributed to this situation, and thereby had a large influence on the design of our Distributed Annotation System for Molecular Interactions (see [Section 3.2](#)), was the unavailability of official software tools for creating or parsing standard-compliant documents within the first years after publication. Given the complexity of the XML document definition, for example, the XML schema definition comprises more than 1000 lines^d, and the con-

^dhttp://psidev.sourceforge.net/molecular_interactions//rel25/src/MIF25.xsd

stantly evolving nature of the MI controlled vocabulary, this initially hampered the usability of PSI-MI.

The PSI-MI standard was developed and released in an incremental approach. While the initial version was only designed for protein interactions, the successor version, HUPO-PSI-MI 2.5^e, extended the scope to molecular interactions in general, also covering protein-ligand interactions or nested protein complexes^{191,271}. The transition phase from PSI-MI 1.0 to version 2.5 spanned over several years, as initially few tools supported the new format³⁷. Fortunately, as of 2013, PSI-MI 2.5 is the widely accepted standard for the representation of molecular interactions.

Arguably a contributing factor for today's wide acceptance was the adaptation of a less complex way for representing interaction data. While PSI-MI XML files allow for a very detailed characterization of experimentally determined molecular interactions, they can hardly be created or parsed without adequate tools. PSI-MI 2.5 now also defines a more lightweight data standard named MITAB2.5, a tab-delimited plain text file format. In the following sections, I will briefly describe the two key components of PSI-MI 2.5, namely, the XML format description and the MITAB format.

2.5.2 PSI-MI 2.5 XML format

Markup languages like XML or the Hyper Text Markup Language (HTML) define document formats that are readable for both humans and computers. While HTML is primarily used to describe the structure of internet websites, XML may represent almost arbitrary data. The structure of an XML documents which is composed of nestable elements with optional attributes and content, is defined by an XML schema (see Section 3.1.2.2). Compared to plain text formats, a great benefit of XML documents is that using such XML schemas, it is possible to validate the content of a document and its elements.

Main elements The PSI-MI 2.5 XML document format is built around three main entities¹⁹¹: `interactor`, `experimentDescription`, and `interaction`. An `interactor` defines the interacting entities, for example, genes, proteins, protein domains, or small molecules. The purpose of this element is not a detailed description of the entity, rather a unique identification via its protein sequence or references to external databases like UniProt³⁷⁵, RefSeq, or Entrez Gene⁴⁰⁶.

An `interactor` element is linked to `interactions` via its role as a `participant`. This way, a once-defined interactor can take part in multiple interactions via different biological or experimental roles, for instance, as bait or as prey in a TAP-MS experiment. The `interaction` element may also contain further information on the interaction itself, including its type or confidence scores.

^eVersion 2.5 directly followed version 1.0, no version 2.0 was officially available.

The `interaction` element not only links to `interactors`, but also to `experimentDescriptions` that specify how an interaction has been determined by providing details on the experimental procedures, detection methods, or protein modifications. As all these elements may occur multiple times, for example, several interactors participate in an interaction that is confirmed by various experiments, they are listed within container elements named `interactorList`, `experimentList`, and `interactionList`, respectively.

While the aforementioned elements are the most important entities of the PSI-MI XML format, there is a multitude of additional elements, which can be used within different contexts to describe aspects like the host organism, details on the level of data curation, or the availability of interaction data. A full description of the XML definition can be found on the PSI-MI website^f, a valid PSI-MI 2.5 XML document describing the binary interaction between two proteins can be found in the Appendix.

Compact and expanded form PSI-MI XML documents may exist in two interchangeable formats. In the compact form, `interactors` and `experimentDescriptions` are defined once in the respective `interactorLists` and `experimentLists` in the beginning of a document. Later occurrences of these elements in an `interaction` are cross-referenced using their `id` attributes.

In the expanded form, `interactorList` and `experimentList` are not defined in the beginning of the PSI-MI document. Instead, each `interactor` and `experimentDescription` is completely defined within the respective `interaction` element. This expanded form has the advantage that each `interaction` element is self-contained, that is, includes all the relevant information without cross-referencing to other parts of the document. This improves document handling, in particular for humans, as jumps within potentially large files are prevented. The downside of this verbosity is an increased file size, as additional markup is required and potentially identical information is repeated multiple times. Both representations are interchangeable and there are tools that allow transforming one representation into the other. The exemplary PSI-MI XML document listed in the Appendix has been downloaded from IntAct in the standard compact form.

2.5.3 MITAB format

MITAB has been defined as a simplified alternative to PSI-MI XML in PSI-MI version 2.5¹⁹¹. The MITAB format was a response to the growing demand within the scientific community for a more lightweight data standard that focuses on reduced file size and, due to less complexity, ease of use. MITAB files are tab-delimited plain text files, each line describes an interaction. The format is based on the BioGRID TAB formats³⁵⁶ and defines the following fifteen mandatory columns:

^f<http://www.psidev.info/mif>

1. Identifier interactor A
2. Identifier interactor B
3. Alternative identifier interactor A
4. Alternative identifier interactor B
5. Alias interactor A
6. Alias interactor B
7. Interaction detection method
8. First author of the publication that first showed the interaction
9. Identifier of the publication
10. NCBI Taxonomy identifier interactor A
11. NCBI Taxonomy identifier interactor B
12. Interaction type
13. Source database
14. Interaction identifier in source database
15. Confidence score

Each column may contain multiple values, which are separated by a vertical dash ("|"). If a column has no entry, this is indicated by a horizontal dash ("-"). Soon after the MITAB format had been released, different databases like IntAct or iRefIndex started to extend it with additional columns to describe data that were not covered in the initial specification. On the basis of the initial MITAB version and these custom extensions, two official format updates have been defined as MITAB2.6²⁷³ and MITAB2.7²⁷¹. MITAB2.6^g extends the 15 columns of MITAB2.5 with 19 additional columns. Among the new columns are descriptions of the experimental and biological roles of the interactors, further external references, taxonomical information of the experimental systems, dates when an interaction record has been created or updated, and checksums that can be used to verify the correctness of the interactors. On the basis of this extension, MITAB2.7^h has been released recently, now defining 42 columns in total. The six new columns in this release can be used to describe general interaction features like binding sites or post-translational modifications, stoichiometry, and participant identification methods for both interactors. MITAB2.7 allows the representation of binary interaction records according to the MIMIx standard²⁸¹. All MITAB extensions are supposed to be backwards compatible.

^g<http://code.google.com/p/psicquic/wiki/MITAB26Format>

^h<http://code.google.com/p/psicquic/wiki/MITAB27Format>

2.5.4 Binary representation of protein complexes

From the MITAB definitions above, it can be seen that the standard was designed for binary interactions that involve one pair of interactors. Protein complexes cannot be represented without different ways of representing the data, as a complex assembly of multiple proteins cannot be split into an interactor A and interactor B. Consequently, different models have been proposed to represent or expand protein complexes in a binary fashion. The three approaches most commonly used are matrix, spoke, and bipartite expansion (see Figure 2.3). In MITAB2.6 and MITAB2.7, the sixteenth column can be used to indicate that a binary entry resulted from a complex expansion.

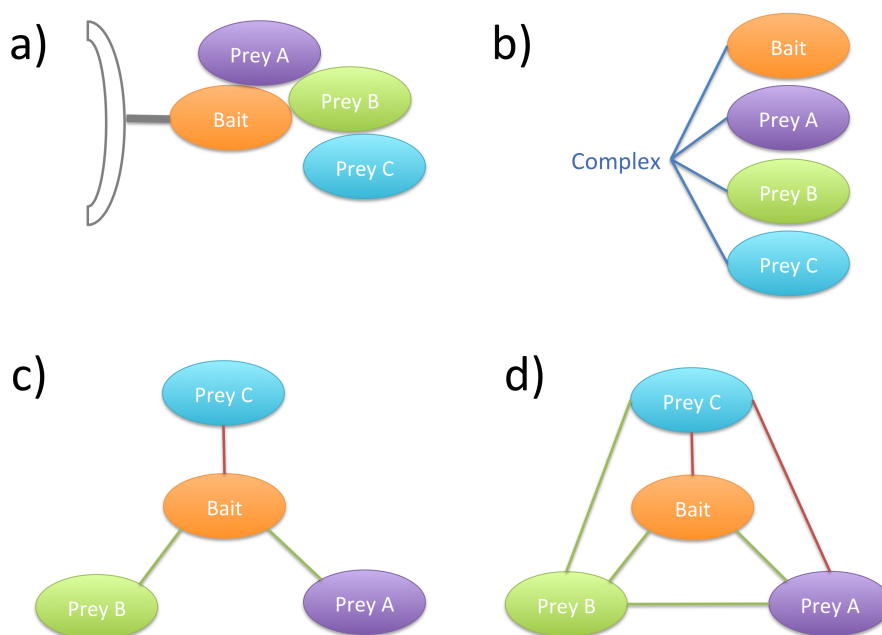


Figure 2.3: Binary representations of protein complexes. **a) Real complex**, exemplarily illustrated with a tagged bait and three preys. Node contacts indicate four binary protein-protein interactions (Bait - Prey A, Bait - Prey B, Prey A - Prey B, Prey B - Prey C). **b) Bipartite representation**, resulting in four virtual binary interactions (blue). **c) Spoke expansion**, resulting in two true positive interactions (green), one false positive (red), and two false negatives (Prey A - Prey B, Prey B - Prey C). **d) Matrix expansion**, resulting in four true positive binary interactions (green) and two false positives (red).

2.5.4.1 Matrix expansion

In a matrix expansion, each protein within a complex is assumed to be interacting with all other proteins apart from itself. The exemplary complex in Figure 2.3, assembled by four proteins, would be represented by six binary interactions. This expansion does not make any assumptions about the actual structure of the complex

and disregards information about baits or preys. Consequently, the number of false negative interactions is zero. However, a number of false positive interactions will be predicted, as not all proposed interactions could take place at the same time in the actual complex¹⁰⁶.

2.5.4.2 Spoke expansion

The spoke expansion of a protein complex is centered on a bait protein, which is used to (virtually) pull out multiple preys. The exemplary protein complex in [Figure 2.3](#), composed of four proteins, would be represented by three binary interactions. Compared to matrix expansion, spoke expansion has a higher number of false negatives, as only interactions between baits and preys are considered and interactions between baits are ignored. The number of false positive and false negative interactions depends on the selected bait. In the illustrated example, choosing "Bait" as the central spoke protein would yield two false negative and one false positive interaction, while selecting "Prey B" as the bait for the spoke expansion would result in one false negative and no false positive interactions.

2.5.4.3 Bipartite representation

Bipartite representation is not an actual complex expansion model, as no real binary protein interactions are deduced. In a bipartite representation, one of the binary interactors is used to represent the complex itself, while the other interactor is used to represent each protein of the complex. The exemplary protein complex in [Figure 2.3](#) would be represented by four binary membership representations. Bipartite representation is used by the aggregate database iRefIndex to prevent the accidental confusion of real binary interactions with expanded complexes in MITAB files³¹⁵.

3 Data decentralization in interactomics

This chapter presents methods that we have developed to simplify the process of working with molecular interaction data, in particular with respect to data accessibility and exchange. [Section 3.1](#), will introduce the Distributed Annotation System (DAS), a decentralized client-server system for genomic and proteomic sequence annotation. This system is the basis for our extension, named DAS for Molecular Interactions (DASMI), which I will describe in detail in [Section 3.2](#). While DASMI has been the first distributed system for molecular interactions, today it is largely superseded by a successor named Proteomics Standards Initiative Common Query Interface that I will briefly describe in [Section 3.3](#). I will end the chapter in [Section 3.4](#) with summarizing remarks and an outlook on potential future developments.

3.1 Introducing the Distributed Annotation System

[Chapter 2](#) highlighted the great importance of molecular interactions for understanding biological and cellular processes and described the resulting growth in the amount of available interaction data. However, it also described, how these data are scattered over a multitude of online repositories, with each providing different means for data access³⁵⁹.

A similar situation existed in the field of genomic sequence annotation around the year 2000, when breakthroughs in sequencing technologies and the competitions for the first sequenced genomes, most notably the one between the International Human Genome Sequencing Consortium and Celera Genomics^{209,389}, led to a substantial increase in available sequence data. In fact, more raw genomic data were produced than the central repositories like GenBank or EMBL could annotate⁹⁴. As a consequence, third parties needed to step in and take over parts of the annotation, for instance, by predicting the position of genes or by computing general sequence features¹⁰⁴. This, however, led to a fragmentation of annotations, which were no longer present in the centralized resources, but spread over a multitude of online repositories and databases.

Dowell et al. developed the Distributed Annotation System (DAS) as one potential solution to overcome this fragmentation⁹⁴. The basic idea of DAS is to replace centralized repositories with distributed storage systems. DAS refers to two things, the general architecture of servers and clients, illustrating the idea of data federation (see Section 3.1.1), and the actual data exchange protocol, which defines the communication between the individual DAS components (see Section 3.1.2). While the DAS architecture is more or less static and has not changed much since its first publication, the DAS protocol is constantly evolving in response to the different demands from the scientific community (see Table 3.1).

As we selected DAS as the basis for our proposal to the challenge of data distribution in interactomics (see Section 3.2), I will describe the individual DAS components in the following sections.

Table 3.1: Versions of the Distributed Annotation System (DAS). The table lists the name, specification URL and if present the year and publication. Versions ending with an E are unofficial protocol versions that bundle extensions.

Version	Year	Specification URL
1.01 ⁹⁴	2001	http://www.biodas.org/documents/spec-1.53.html
1.53	2002	http://www.biodas.org/documents/spec-1.53.html
1.53E ¹⁸¹	2008	http://www.dasregistry.org/spec_1.53E.jsp
1.6	2010	http://www.biodas.org/documents/spec-1.6.html
1.6E	-	http://www.biodas.org/wiki/DAS1.6E
2.0	2006	http://biodas.org/documents/das2/das2_protocol.html

3.1.1 DAS architecture

DAS is designed as a client-server system, where servers provide different types of biological information that clients integrate and visualize for the end user. DAS distinguishes two types of servers. Reference servers (S_R) provide the biological reference, commonly a nucleotide or protein sequence. Annotation servers (S_A) make additional information available, which can be related to the reference entity, for example, the position of exons, introns, or protein domains.

In the most basic form, DAS can be classified as a $CS_R S_A^+$ system, indicating that one client (C) visualizes the reference entity provided by a single reference server along with annotations from at least one annotation server. In modern practice, DAS clients may embed other clients that incorporate a different reference entity, for instance, the DAS-based Ensembl genome browser that incorporates a protein structure view³⁰¹.

The distinction between reference and annotation server does not have to be physical. Depending on the query type, a single DAS server instance can act as

both, reference and annotation server. In the DAS terminology, there is a further distinction between DAS servers and sources. The term server is used for the actual server instance, which is set up as a specific piece of software on a (physical or virtual) server machine. Each dataset that is provided by such a server is referred to as a source.

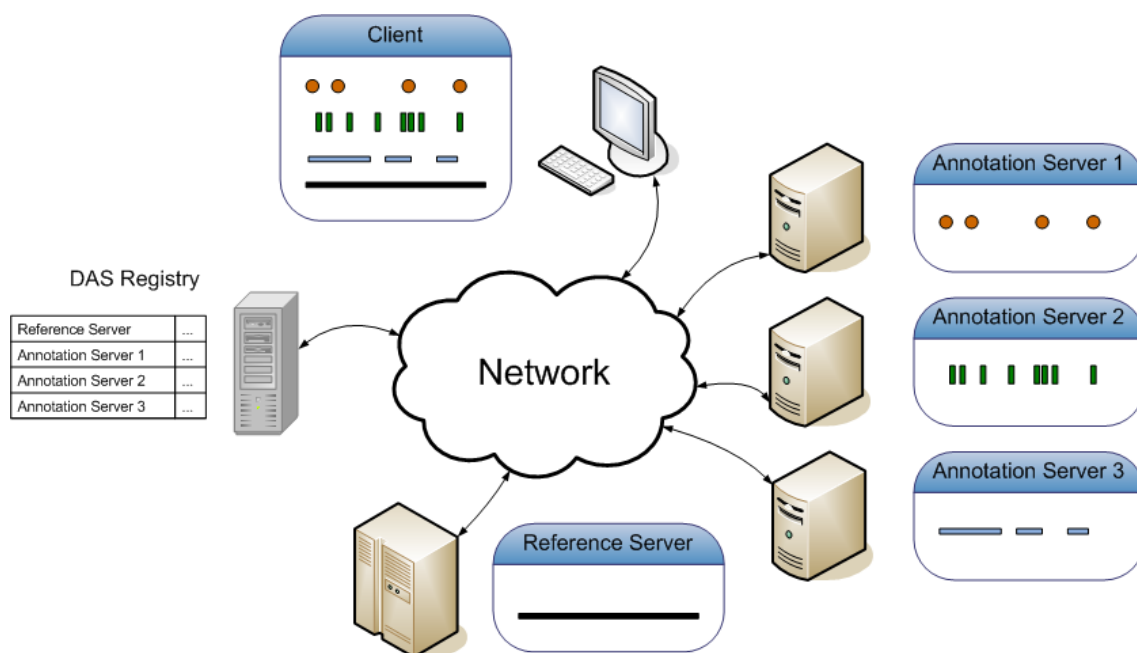


Figure 3.1: Schematic drawing of the client-server architecture of the Distributed Annotation System (DAS)^{37,94}. Reference servers provide the reference sequence, annotation servers one to three make additional information available, which can be related to this reference entity. A DAS client communicates with all servers it retrieves from the DAS registry, unifies the individual results, and presents them to the user.

3.1.1.1 Reference server

Reference servers provide the biological entity that annotations are relating to. In the initial DAS implementation, the reference was a nucleotide sequence, for instance, a chromosome or a gene⁹⁴. As DAS evolved, proteomics became a second large domain, with the reference entity being the amino acid sequence of a protein²⁶⁷. For those two areas, the reference sequence is usually curated and provided by large consortia like Ensembl¹⁶⁷ or UniProt^{12,375}. In the context of the eFamily project, Prlić et al. have extended the use of DAS beyond the sequence world, with a DAS extension that uses protein structures as reference entities^{299–301}. Within the BioSapiens framework, DAS was further used for the exchange of proteomics data in general, not limited to protein sequences and their annotations³⁷⁷. The biological entity type, which a server is using, is specified with specific a *coordinate system* (see Section 3.1.1.5).

3.1.1.2 *Annotation server*

The majority of DAS servers are annotation servers, which provide different types of information that can be related to the reference entity (see [Figure 3.1](#)). The information is either positional, such as the coordinates of genes, protein domains, introns, exons, signal peptides, or splice sites, or non-positional, such as links to scientific publications or database identifiers. Like reference servers, annotation servers are assigned to certain coordinate systems that specify the biological entity their annotations are relating to, for instance, genes, proteins, or chromosomes (see [Section 3.1.1.5](#)).

In the early days of DAS, there was little regulation of the annotation features that a server provided, that is, features were described in free text, with a few proposals listed in the official specification⁹⁴. While this promoted a rapid adaptation of DAS, as setting up servers and providing sequence features was straightforward, it also limited the analyses capabilities, as automated exploration of the features was hampered by the free text annotations, which were more targeted towards humans than computers. In order to improve computational analysis of DAS annotations, for instance, allowing a grouping of similar features, an ontology of feature types has been developed within the BioSapiens framework^{181,317}.

3.1.1.3 *Client*

One of the fundamental principles of DAS, is the separation of data, which are offered by the two types of DAS servers, from their visualization that is performed by DAS clients. As DAS has its origins in sequence annotation, most of the existing DAS clients are genome or proteome browsers, tools that display sections of nucleotide or amino acid sequences alongside user-selected feature tracks. These browsers do not necessarily retrieve all their data from DAS servers, but combine DAS results with other data sources, for instance, from centralized data warehouses¹⁶⁷. Generally, available DAS clients can either be accessed through a website, such as Ensembl¹⁶⁷, Genome Maps²⁴², Dalliance⁹⁵, the CBS viewer²⁶⁷, MyKaryoView³⁹³, and Dasty^{185,393}, or run on local computers as stand-alone tools, such as the Integrated Genome Browser (IGB)¹⁵⁰, Spice³⁰⁰, and Jalview⁷³. A feature almost all modern DAS clients support, is the communication with the DAS registry in order to determine the available DAS sources and allow the user to select the desired ones.

3.1.1.4 *Registry*

In the original implementation, DAS did not include a component for keeping track of the, initially few, available servers and sources. As a result, in order to incorporate a DAS source, clients or their human end users needed to know its URL. The steady growth in both, the number of available DAS sources and their diversity in terms of provided features, necessitated the development of a central repository that maintains all these data in a format accessible for humans and DAS clients. The

resulting service, named DAS registry, was developed at the Wellcome Trust Sanger Institute by Prlić et al.³⁰¹. Service providers may register their DAS sources to make them available to all DAS clients that communicate with the registry. In addition, the registry constantly checks the availability of every server and tests their conformity with the DAS specification. As of November 2013, the DAS registry^a lists more than 1 500 DAS sources from 18 different countries.

3.1.1.5 Coordinate system

Coordinate systems define the biological entities that DAS sources are providing or annotating, for example, human chromosomes, protein structures, or UniProt proteins. Coordinate systems have been introduced into DAS with the appearance of the DAS registry, as means for automatically classifying the available data³⁰¹. Each coordinate system is composed of an authority (the responsible institution, for example, NCBI, Ensembl, or UniProt), an optional version, an optional organism taxonomy identifier, and a type (for example, protein sequence, chromosome, or protein structure). Further, it is identified via a unique Uniform Resource Identifier (URI), currently based on the Uniform Resource Locator (URL) of the DAS registry. Two coordinate systems, one based on NCBI Human genome assembly 36 and one based on UniProt sequences, are shown in the XML document in Listing 3.1. All coordinated systems known to the DAS registry, 791 as of November 2013, can be requested via a specific URL^b.

```
1 <DASCOORDINATESYSTEM>
2   <COORDINATES uri="http://www.dasregistry.org/dasregistry/coordsys/CS_DS40"
3     taxid="9606" source="Chromosome" authority="NCBI" test_range="" version
4     ="36">
5     NCBI_36,Chromosome,Homo sapiens
6   </COORDINATES>
7   <COORDINATES uri="http://www.dasregistry.org/dasregistry/coordsys/CS_DS6"
8     source="Protein Sequence" authority="UniProt" test_range="P15498">
9     UniProt,Protein Sequence
10 </COORDINATES>
</DASCOORDINATESYSTEM>
```

Listing 3.1: XML representation of two DAS coordinate systems

3.1.2 DAS protocol

The DAS protocol defines the communication between DAS servers and clients. In contrast to the general architecture (see Section 3.1.1), which, apart from the introduction of a registry, remained more or less unchanged of the years, the DAS protocol repeatedly evolved by introducing new application areas (see Table 3.1).

^a<http://www.dasregistry.org/>

^b<http://www.dasregistry.org/das/coordinatesystem>

The initial protocol version 1.01 was tailored towards the annotation of nucleotide sequences⁹⁴. Version 1.53, released shortly after in 2002, brought improvements like error handling. Importantly, it extended DAS for protein sequence annotation by introducing `sequence` as a general alternative to the nucleotide-specific `dna` command (see [Section 3.1.2.1](#)). This version of DAS became increasingly popular and was the standard for several years. As it was also the basis for our molecular interaction extension to the DAS system (see [Section 3.2](#)), all references to DAS in the remainder of this chapter will refer to DAS1.53, unless otherwise stated.

The DAS protocol is deliberately simple. Requests have the form of Hypertext Transfer Protocol (HTTP) Uniform Resource Locators (URLs), responses are XML documents of limited complexity. The intention behind this was to maximize the support in the biological community, by making the whole protocol human readable and easily parseable⁹⁴. The downside of this simplicity is limited expressiveness and, due to the verbose XML response formatting, a potentially large overhead in document size.

DAS/2 The fixation of DAS on verbose XML was one of the main reasons, why in 2004 work on DAS/2 started. In contrast to all 1.x versions, DAS/2 was not intended to be backwards compatible. It was designed as a completely new protocol, with the aim of making the whole system more powerful and flexible. However, after an ambitious and promising start, backed by an National Institutes of Health (NIH) grant from 2004 to 2006, work on DAS/2 has stagnated over the last years and there is very little ongoing development. As of 2013, only the software suite *GenoViz*¹⁴⁹ and the IGB client built upon it¹⁵⁰ support DAS/2. Consequently, there is no support in the DAS registry, and only a few server instances are currently available.

Extensions Especially in Europe, since the mid 2010th the driving force of DAS development^{95,117,132,181,185,259,267,300,301,329,330,387}, where most of the currently available DAS servers are hosted, there was a strong opposition towards a backward-incompatible DAS version. The approach favored by the European DAS community is to extract the parts from DAS/2 that could improve the overall usability and incorporate them into the DAS standard in the form of extensions. This was pioneered with DAS1.53E¹⁸¹, an unofficial DAS version that combines DAS1.53 with three DAS extensions, one of them being our molecular interaction extension that is described in [Section 3.2](#).

If extensions are well accepted by the community and are widely used, they may be incorporated into a later version of the official DAS standard. This was shown by DAS1.6, the latest official protocol specification, which was released in 2010. In addition to general clarifications and cleanups, this version officially included support for coordinate systems and the DAS registry. Since the release of DAS1.6, there have been a number of additional extensions and proposals, such as write-back, that is, the possibility to upload comments or changes from the client to the server,

search, or authentication capabilities. These are bundled into an unofficial version DAS1.6E. Table 3.2 lists all DAS requests and associated response formats of the different official DAS version.

Table 3.2: DAS commands and associated XML response formats from the different official DAS specifications (see also Table 3.1). Unofficial commands from the extended specifications 1.53E and 1.6E are not listed. The commands `dsn`, `dna`, and `link` have been deprecated in version 1.6. S_A indicates that a command can be issued to an annotation server, S_R marks commands suitable for reference servers.

Command	Server type	DAS versions	XML response	Description
<code>dna</code>	S_R	1.01-1.53	DASDNA	Nucleotide sequence of the corresponding segment.
<code>dsn</code>	S_A, S_R	1.01-1.53	DASDSN	Available data sources
<code>entry_points</code>	S_R	1.01-1.6	DASEP	Either entry points into a sequence (e.g. chromosomes) or reference objects (e.g. proteins)
<code>features</code>	S_A, S_R	1.01-1.6	DASGFF	Annotations for a reference object
<code>link</code>	S_A, S_R	1.01-1.53	-	Additional human-readable information on annotations, e.g., a web page
<code>sequence</code>	S_R	1.53-1.6	DASSEQUENCE	Nucleotide or protein sequence of the corresponding segment
<code>stylesheet</code>	S_A, S_R	1.01-1.6	DASSTYLE	Server recommendations for displaying annotations
<code>structure</code>	S_R	1.6	DASSTRUCTURE	Three-dimensional structural coordinates
<code>types</code>	S_A, S_R	1.01-1.6	DASTYPES	Available annotation types

3.1.2.1 DAS requests

Requests to a DAS server are issued in the form of a Hypertext Transfer Protocol (HTTP) Uniform Resource Locator (URL). These URLs have a specific format that is composed of:

- a site-specific prefix (`http://www.ebi.ac.uk/das-srv/uniprot/`),
- the keyword `das`,

- a data source name (`uniprot`),
- a DAS command (`sequence`),
- optional parameters, depending on the DAS command (`segment=P09497:0,50`).

Joined by the appropriate HTTP connectors, the above example^c would retrieve the first 50 amino acids of the Clathrin light chain B protein (UniProt accession number P09497) from the UniProt DAS server that is hosted at the European Bioinformatics Institute (EBI). Table 3.2 lists all commands that are part of official DAS specifications, commands that are part of unofficial extensions like `alignment`³⁰¹, `volmap`²²⁹, or `interaction` (see Section 3.2.1.1) are not listed. Most DAS commands accept additional parameters to further specify a query, for example, by selecting a certain range or feature types. These parameters and some exemplary queries can be found in the respective DAS specifications, which are listed in Table 3.1.

3.1.2.2 DAS response

A DAS server responds to a request with an XML document that is wrapped in an HTTP envelope. Since DAS1.53, this envelope does not only contain standard HTTP header information, but additional DAS status codes, which can be utilized by DAS clients. Each of the DAS commands is associated with a server response that is formatted according to a particular XML schema (see Table 3.2). Listing 3.2 shows the response to the aforementioned exemplary `sequence` request. As can be seen, the response document is very simple, only consisting of a `DASSEQUENCE` root element that contains one `SEQUENCE` element. This `SEQUENCE` element has the attributes `id`, `start`, `stop`, `version`, and `label` and contains the actual amino acid sequence in its main element content.

```
1 <DASSEQUENCE>
2   <SEQUENCE id="P09497" start="1" stop="50" version="9484
3     bcf078b4b32f17f823a308d3ba5" label="Clathrin light chain B">
4     MADDGFFSSSESGAPEAAEEDPAAAFQAQSESEIAGIENDEGFGAPAGS
5   </SEQUENCE>
</DASSEQUENCE>
```

Listing 3.2: Exemplary `DASSEQUENCE` XML document.

XML schema definition A great strength of XML and other markup languages is that their structure can be defined via a schema. Thus allows testing documents for their standard conformity and validity. There are various languages for defining an XML schema; this section will only briefly introduce the three that are relevant for DAS.

^c<http://www.ebi.ac.uk/das-srv/uniprot/das/uniprot/sequence?segment=P09497:0,50>

Document Type Definitions (DTDs) are one of the oldest ways to specify markup languages. While they allow for a very compact and simple definition, the limited expressiveness, mostly due to the missing support for namespaces and other elements than strings, were major reasons why two alternative schema languages have been implemented around the year 2001.

XML Schema Definitions (XSDs) are the official recommendation of the World Wide Web Consortium (W3C) for defining the content of XML documents, as such they have the largest support in the community. The REgular LAnguage for XML Next Generation (RELAX NG) is an alternative schema language that appeared around the same time and has many similarities to XML Schema. Both languages are themselves based on XML and can thus make use of namespace and complex element types (see [Listing 3.4](#) and [Listing 3.5](#)).

Unfortunately, there is a mixture of schema definitions in the different DAS versions. In the first versions, 1.01 and 1.53, all documents were defined via DTDs⁹⁴. The extensions to DAS1.53 were defined using XML Schema Definitions, leading to a mixture of DTDs and XSDs in DAS1.53E¹⁸¹. DAS/2, the re-design of DAS1, based their schemas on RELAX NG. DAS1.6, the latest official version, subsequently also replaced their DTD schemas with RELAX NG specifications.

In order to compare the three schema definition languages, listings 3.3, 3.4, and 3.5 show the corresponding definitions of the DASSEQUENCE XML format, a valid XML document is shown in [Listing 3.2](#). The slight discrepancies between the three schemas are due to the different DAS versions they are based on. For example, the attribute `molecule` that is defined in the DTD was deprecated in DAS1.6 and a new, optional `label` attribute was introduced.

```
1 <!ELEMENT DASSEQUENCE (SEQUENCE+)>
2 <!ELEMENT SEQUENCE (#PCDATA)>
3 <!ATTLIST SEQUENCE id CDATA #REQUIRED>
4 <!ATTLIST SEQUENCE start CDATA #REQUIRED>
5 <!ATTLIST SEQUENCE stop CDATA #REQUIRED>
6 <!ATTLIST SEQUENCE version CDATA #REQUIRED>
7 <!ATTLIST SEQUENCE molecule CDATA #REQUIRED>
```

Listing 3.3: DASSEQUENCE schema as Document Type Definition (DTD)

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
3   elementFormDefault="qualified">
4   <xs:element name="DASSEQUENCE">
5     <xs:complexType>
6       <xs:sequence>
7         <xs:element maxOccurs="unbounded" ref="SEQUENCE"/>
8       </xs:sequence>
9     </xs:complexType>
10  </xs:element>
11  <xs:element name="SEQUENCE">
12    <xs:complexType mixed="true">
```

```

13     <xs:attribute name="id" use="required"/>
14     <xs:attribute name="start" use="required" type="xs:integer"/>
15     <xs:attribute name="stop" use="required" type="xs:integer"/>
16     <xs:attribute name="version" use="required"/>
17   </xs:complexType>
18 </xs:element>
19 </xs:schema>

```

Listing 3.4: DASSEQUENCE schema as XML Schema (XSD)

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <grammar xmlns="http://relaxng.org/ns/structure/1.0" xmlns:a="http://relaxng.org/
   ns/compatibility/annotations/1.0" datatypeLibrary="http://www.w3.org/2001/
   XMLSchema-datatypes">
3   <define name="DASSEQUENCE">
4     <element name="DASSEQUENCE">
5       <oneOrMore>
6         <element name="SEQUENCE">
7           <attribute name="id"><text/></attribute>
8           <attribute name="start"><data type="integer"/></attribute>
9           <attribute name="stop"><data type="integer"/></attribute>
10          <optional><attribute name="version"><text/></attribute></optional>
11          <optional><attribute name="label"><text/></attribute></optional>
12          <text/>
13        </element>
14      </oneOrMore>
15    </element>
16  </define>
17  <start>
18    <ref name="DASSEQUENCE"/>
19  </start>
20 </grammar>

```

Listing 3.5: DASSEQUENCE schema as RELAX NG

3.1.2.3 Limitations of DAS

Since its publication in 2001, DAS has seen a substantial growth in popularity¹⁸¹. It has been used as an architecture for proteome annotation within European projects like BioSapiens and eFamily and is supported by major bioinformatics resources like the Ensembl genome browser or the UniProt knowledge base^{192,299,317}. As of November 2013, there are more than 1 500 DAS sources from eighteen countries registered in the DAS registry.

Despite this success, DAS still has a number of limitations, several of those have already been discussed in my Diploma thesis six years ago³⁷. Still amongst them is the missing access control, as DAS does not specify any user authentication. Everybody knowing the URL of a DAS source can access its content, which is problematic especially with respect to personalized genomic data. Another weakness is the absence of a search functionality, that is, the ability to determine if a certain feature is

present in a server, without retrieving all its content. The large footprint of a DAS document, due to the use of verbose XML, and the lacking support for binary data, limits the use of DAS for the annotation of next-generation sequencing data.

The DAS community has acknowledged many of the limitations. A particular forum for discussions were the annual workshops at the EBI, which were held between 2008 to 2012. While there are proposals in DAS/2 and DAS1.6E for user authentication and search functionality, no robust implementation is yet available.

3.2 DAS for Molecular Interactions

Chapter 2 described the reasons, why interaction data were being made available in a multitude of online repositories, hampering easy access for researchers. One way of trying to improve this situation was shown by aggregate interaction databases like iRefIndex or STRING, which collect and combine data from a number of primary resources (see Section 2.3.2). However, the static data integration performed by those aggregate databases has several drawbacks, as it only provides a fixed snapshot of a number of databases at a certain point of time. Aggregate resources can only remain relevant for research if they ensure that their data is still more or less in sync with the original sources, which are often continuously updated⁵⁸. In addition, the authority that hosts the aggregate resource usually performs the integration of data centrally. As a result, the data cannot easily be changed or extended by the user, for instance, with confidential results from in-house experiments.

When we started working on a different approach for interaction data aggregation in 2006, our aim was to overcome the aforementioned shortcomings of the static integration frameworks by adopting a decentralization strategy like it was shown by the Distributed Annotation system (DAS, see Section 3.1). DAS comprises a number of servers, providing the relevant data, and clients that integrate those data into a unified view for the user. The standardized protocol, which clients and servers use for communication, and the availability of server software, allows for easy extensibility of the system. In contrast to aggregate databases, there is theoretically no additional maintenance effort, as the data are kept with their original producers and are therefore automatically up to date.

Our system, named DAS for Molecular Interactions (DASMI), consists of three main components (see Figure 3.2):

1. the specification of a DAS extension for exchanging interaction data and additional annotations,
2. DAS servers for providing interaction data using the extended DAS protocol,
3. DAS clients to unify and visualize the interaction data in a suitable way.

A prototype of DASMI has been defined during the course of my Diploma thesis³⁷. Building on this prototype, the first official DASMI version was described in

a dedicated publication⁴⁰ and, together with other extensions to the DAS protocol, in the DAS1.53E bundle¹⁸¹. The DASMIweb client (see Section 3.2.3.1) has further been described separately as a gateway to interactome data⁴¹. In the following, I will summarize the findings from these publications and introduce the different DASMI components in detail.

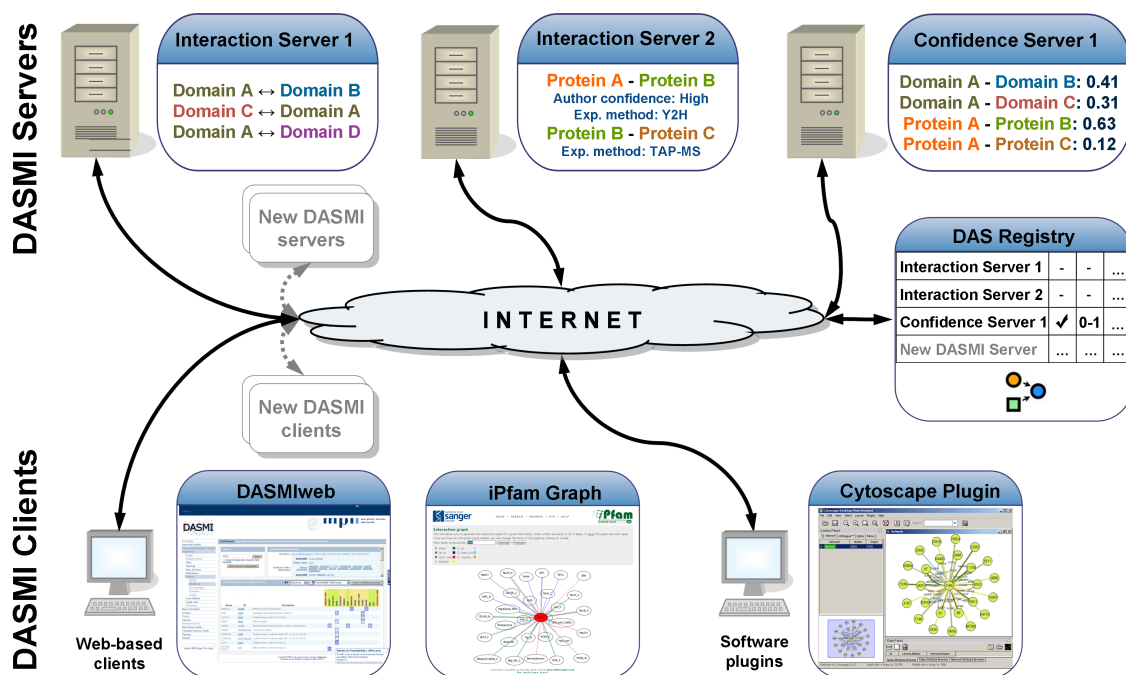


Figure 3.2: Schematic drawing of the client-server architecture of the Distributed Annotation System for Molecular Interactions (DASMI)⁴⁰.

3.2.1 DAS protocol extension

Our DAS protocol extension defines, how molecular interaction data and additional annotations are exchanged between servers and clients. At the project initiation in 2006, we faced the question, whether an extension to the DAS protocol, that is, the definition of a new request command and associated response format, was actually required, or if the data could also be transferred using the protocol that was already in place.

Protocol overloading An example of protocol overloading was provided by the DAS server that had been set up at the Max Planck Institute for Informatics (MPII) as part of our contribution to the ENCODE protein complement annotation³⁷⁹. The server^d used the normal `features` command for interaction requests. Consequently, it overloaded the associated `DASGFF` response format in order to transfer protein

^dThe server is no longer available as of November 2013

interaction data (see Listing 3.6). Overloading here means that the `METHOD` element, which, according to the DAS specification, should describe the method that has been used for detecting a certain feature, was used to identify the interaction partner of a query protein. Another proposal for DAS document overloading, provided by Antony Quinn in a personal communication, was to use the `TARGET` element to identify the interaction partner and use `METHOD` to describe the experimental interaction detection technique.

Both approaches use XML elements in ways that are not according to the DAS specification. This can easily lead to problems with DAS clients that strictly follow the specification. In addition, the expressiveness of the overloaded documents is rather limited, as all the information has to be fitted into the few elements of the DASGFF format. Complex cases, like the interaction between two protein complexes that are each assembled by various proteins, would only fit into such a schema using a number of tricks.

```
1 <SEGMENT id="RP4-696P19.1-002" start="1" stop="1" version="1.0">
2   <FEATURE id="Q5VU07:RP4-696P19.1-002/295189" label="RP4-696P19.1-002">
3     <TYPE id="Q5VU07:PPI-exp:VIDAL" category="miscellaneous">Q5VU07:PPI-exp:VIDAL
4     </TYPE>
5     <METHOD id="Q5VU07">Q5VU07</METHOD>
6     <START>0</START>
7     <END>0</END>
8     <SCORE>-</SCORE>
9     <ORIENTATION>+</ORIENTATION>
10    <PHASE>0</PHASE>
11    <LINK href="http://www.ebi.uniprot.org/uniprot-srv/uniProtView.do?proteinAc=
12    Q5VU07">Q5VU07</LINK>
13    <NOTE>PPI source: VIDAL, Interaction partner:UNIPROT:Q5VU07, PUBMED
14    :16189514</NOTE>
15  </FEATURE>
16 </SEGMENT>
```

Listing 3.6: DASGFF feature representation of a binary protein-protein interaction

Protocol extension Faced with a similar question, Prlić *et al.* chose an alternative approach, when searching for a way to utilize DAS for the exchange of protein structures and their alignments. They developed two new DAS request commands, `alignment` and `structure`, and the corresponding response formats `DASALIGNMENT` and `DASSTRUCTURE`³⁰¹. In addition, they developed the client software SPICE, as no DAS client was capable of displaying protein structures³⁰⁰. Another extension to the DAS protocol was proposed by Macías *et al.*, who developed the `volmap` command and a DAS client named PeppereR to distribute, annotate, and visualize volume maps created with electron microscopy²²⁹. Compared to protocol overloading, the design of a protocol extension has the benefit that it can be fully tailored towards its designated use-case, maximizing the expressiveness of the format. Importantly, an extension only affects DAS infrastructure that supports the extension, legacy

infrastructure, which has been developed for classic DAS commands, can operate unchanged.

The design of our molecular interaction extension was strongly influenced by the extensions of Prlić *et al.* In close collaboration with its main author, Andreas Prlić, Robert Finn, and Andrew Jenkinson, all at the time working at the Wellcome Trust Sanger Institute or the European Bioinformatics Institute, in Hinxton, UK, we defined a new request command named `interaction` (see Section 3.2.1.1) and the corresponding DASINT response format (see Section 3.2.1.2)¹⁸¹.

3.2.1.1 Request: DAS *interaction* command

Requests to a DASMI server are issued in the same form as to a standard DAS server, as an HTTP URL (see Section 3.1.2.1). We defined the new command named `interaction`, which accepts three additional parameters: `interactor`, `detail`, and `operation`. The syntax of the `interaction` command is shown in Listing 3.7. The following description is largely based on the initial definition of the `interaction` command in⁴⁰.

```

1 PREFIX/das/DSN/interaction?interactor=AN_ID
2   [;interactor=AN_ID]
3   [;operation=INTERSECTION_UNION}]
4   [;detail=[ref:AN_ID,]property:A_PROPERTY]
5   [;detail=[ref:AN_ID,]property:A_PROPERTY,value:A_VALUE]
```

Listing 3.7: Syntax of the `interaction` DAS command

The `interactor` parameter represents an identifier of the query interactor. One or several `interactor` parameters can be provided. If only one `interactor` parameter is present, all available interaction partners of this interactor will be returned. The optional `operation` parameter can be used for defining the response type if more than one `interactor` parameter is provided. If `operation` is not provided or has the value "intersection", the interaction partners that interact with all query interactors will be returned. If `operation` has the value "union", the interactions for each of the query interactors will be returned.

The `detail` parameter can be used to refine a query further. If the optional `ref` parameter is used, it indicates that the `detail`, for example, an experimental role, is linked to the interactor with the same ID. If `ref` is omitted, the `detail`, for example a literature reference, is linked to the interaction. A `detail` can be described in a short and a long form. The short form `detail=property:A_PROPERTY` will return all interactions for which there is a certain `property` of the type `A_PROPERTY`, regardless of what the value of this property is. The long form `detail=property:A_PROPERTY, value:A_VALUE` will return all interactions for which a certain `property` of the type `A_PROPERTY` is present and has a specific value (`A_VALUE`). Examples for the short form are the retrieval of all interactions that contain a confidence score or a literature reference, whereas the long form would retrieve only those interactions for which the

confidence score or literature reference has a specific value.

Examples In the following exemplary queries, PREFIX represents the base URL to a DAS server (e.g. <http://dasmi.de/>), followed by the keyword `das`, and DSN, which indicates a particular data source name (e.g. `intact`).

- PREFIX/das/DSN/interaction?interactor=3406
Request all interactions for the interactor with the identifier 3406, in this case representing an Entrez Gene ID.
- PREFIX/das/DSN/interaction?interactor=PF00121
&detail=property:confidenceScore
Request all interactions for the protein domain with the Pfam identifier PF00121, which have an assigned confidence score, regardless of the value of that score.
- PREFIX/das/DSN/interaction?interactor=P52597
&interactor=P60010
Request all interactions that involve the proteins with the UniProt accession numbers P52597 and P60010.
- PREFIX/das/DSN/interaction?interactor=P52597
&interactor=P60010&operation=UNION
Request interactions, in which either the interactors with the UniProt accession numbers P52597 or P60010 are participating.
- PREFIX/das/DSN/interaction?interactor=P38330
&interactor=P32324&interactor=P10591&interactor=P16140
Request protein complexes, in which the four proteins with the UniProt accession numbers P38330, P32324, P10591, and P16140 are participating.
- PREFIX/das/DSN/interaction?interactor=P38330
&detail=ref:P38330,property:experimentalRole,value:bait
Request all interactions, in which the interactor with the UniProt accession number P38330 has the experimental role bait.

3.2.1.2 Response: *DASINT XML format*

This section defines the response to a DASMI request. Section 3.2.1 described, that in order to prevent confusion and maximize the expressiveness, we decided against overloading the existing DASGFF format.

Alternative formats Instead of developing a wholly new XML format, one option would have been to wrap an existing HUPO-PSI-MI format within a DAS envelope. When we started this project in 2006, PSI-MI XML was available in version 1.0¹⁵³. In Section 2.5.2, I have described PSI-MI XML as a very detailed and complex data

format with numerous elements in a deeply branched hierarchy, hardly manageable without adequate software tool support. This can also be seen in the exemplary PSI-MI file listed in Appendix 1, which describes a single binary interaction using more than 250 lines of XML code. PSI-MI XML and our data format DASINT can thus be regarded as complementary approaches. PSI-MI XML is used to exchange comprehensive and static interaction datasets between databases, whereas DASINT is used to provide a fast and dynamic overview with entry points to other databases for further studies. In that sense, DASINT follows the basic principle of DAS, as it aims to be as simple as possible without compromising expressiveness.

After its publication in 2007, the MITAB2.5 format¹⁹¹ would have been another possible data exchange format (see Section 2.5.3). However, several reasons spoke against adapting this format. First, by the time the new standard appeared, we had already implemented a working prototype based on DASINT. Second, MITAB2.5 cannot be used to represent protein complexes without transforming the data into a representation like the spoke or matrix model (see Section 2.5.4). Third, unlike all other DAS formats, MITAB is a tab-delimited flat file format and not an XML format, which would have resulted in acceptance problems with the DAS community.

In summary, at the time neither PSI-MI XML nor MITAB were suitable to act as response formats in our DAS extension. Consequently, we developed the DASINT format that is present in the following.

XML schema Due to several optional elements and attributes, the length of the DASINT XML format may vary significantly. Figure 3.3 shows an illustration of the DASINT schema definition, the complete W3C XML Schema Definition can be found in Appendix 1 or the Supplemental data of Blankenburg et al.⁴⁰

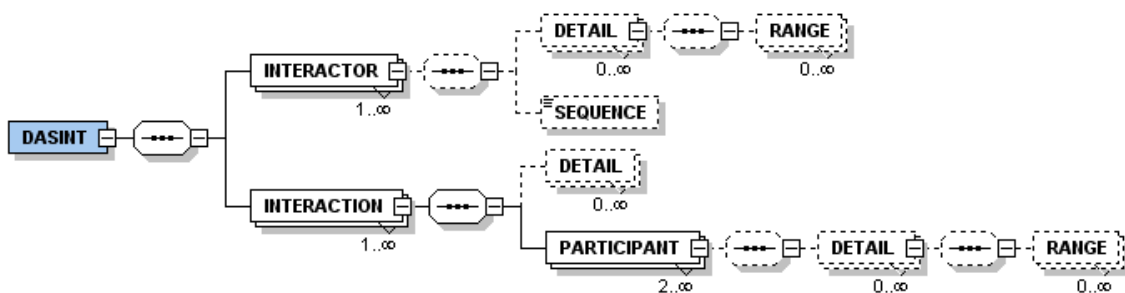


Figure 3.3: DASINT XML Schema Definition.

The main XML elements within the root DASINT element are called INTERACTOR, INTERACTION, PARTICIPANT, and DETAIL. To avoid redundancy and reduce the document size, we decided to use cross-references within the XML document, inspired by the compact form of the PSI-MI XML format (see Section 2.5.2). Cross-referencing means that interacting molecules are defined only once (INTERACTOR) and later participations of these molecules (PARTICIPANT) in interactions (INTERACTION) are linked to the initial definition.

The `INTERACTOR` element contains general, interaction independent information. Mandatory attributes are `shortLabel`, `dbSource`, `dbAccessionId`, and `dbCoordSys`, which define the source of the interactor. The optional `dbSourceCvId` and `dbVersion` can be used to define the source database in more detail. The mandatory `INTERACTOR` attribute `intId` is used to link interactors to their participations in interactions. The sequence of an interactor may be contained in the `SEQUENCE` element. The type of sequence, nucleotide or amino acid, is implicitly defined by the coordinate system of the interactor.

The `INTERACTION` element has the mandatory attributes `shortLabel`, `dbSource`, and `dbAccessionId` that define the origin of an interaction. As for the `INTERACTOR`, the attributes `dbSourceCvId` and `dbVersion` can be used to define a source database in more detail. The sole attribute of the `PARTICIPANT` element is `intId`, which is used to link this element to its initial `INTERACTOR` definition.

The long-term sustainability of the protocol should be ensured by adopting a `property/value` attribute combination for the `DETAIL` element. This enables transferring all different kinds of information, even the kind of information not specified today. For example, including a `DETAIL` element with the attribute "direction" and one of the values "in" or "out" within `PARTICIPANT` elements, allows for assigning directions to interactions and thus for representing pathways in the `DASINT` format. Optional `RANGE` elements may be used to describe positional details like the boundaries of protein domains or binding sites. To specify the provided information in terms of controlled vocabularies, we introduced several optional attributes ending with the term `CvId`.

Listing 3.8 shows an exemplary XML response that contains two proteins and one interaction between them. The interaction is annotated with additional information in the form functional similarity based confidence scores³⁴².

```

1 <DASINT xsi:schemaLocation="http://dasmi.de/dasint.xsd">
2   <INTERACTOR intId="1" shortLabel="HNRPF_HUMAN" dbSource="uniprotkb"
      dbSourceCvId="MI:0486" dbVersion="13.3" dbAccessionId="P52597" dbCoordSys="
      UniProt,Protein Sequence"/>
3   <INTERACTOR intId="2" shortLabel="ACT_YEAST" dbSource="uniprotkb" dbSourceCvId
      ="MI:0486" dbVersion="13.3" dbAccessionId="P60010" dbCoordSys="UniProt,
      Protein Sequence"/>
4   <INTERACTION name="P52597-P60010" dbSource="FunSimMat" dbVersion="2.0"
      dbAccessionId="">
5     <DETAIL property="BPscore" value="0.976656009021968"/>
6     <DETAIL property="CCscore" value="0.488117560317015"/>
7     <DETAIL property="MFscore" value="0.463817012737284"/>
8     <PARTICIPANT intId="1"/>
9     <PARTICIPANT intId="2"/>
10  </INTERACTION>
11 </DASINT>

```

Listing 3.8: Exemplary `DASINT` XML document describing a binary protein-protein interaction with additional confidence scores

3.2.2 DASMI server

A DASMI server processes interaction requests by providing interaction data and potential annotations in the DASINT XML format. As in the original DAS architecture, which consists of reference and annotation servers, there are different types of servers in DASMI (see [Figure 3.2](#)). The majority of DASMI servers provide interaction data and optionally additional information, making them the equivalent of reference servers. Examples for additional information, which can be related to particular interactions, are predicted or known interaction regions or the conditions under which the interaction occurs. Confidence servers, the equivalent of annotation servers, provide reliability scores for interactions.

Compared to the original DAS, the distinction between the two server types in DASMI is not made via different DAS commands (see [Table 3.2](#)). In DASMI, both server types respond to the same `interaction` command with a DASINT XML format, making them both capable of acting as reference and annotation servers. The distinction is made via additional server annotation in the form of keywords, which a DASMI client may use to incorporate those servers in a different way (see [Section 4.2.1](#)).

Each DASMI server belongs to a coordinate system (see [Section 3.1.1.5](#)) that specifies how interaction data can be retrieved from the server, for example, if it uses UniProt accession numbers or Entrez Gene identifiers for describing the interactions.

3.2.2.1 DAS server software

The DAS community has developed several software tools to aid in the process of setting up a DAS server. The most commonly used server tools are ProServer¹¹⁷, implemented in Perl, and the Java-based Dazzle¹³⁴ and MyDAS³²⁹. During my Diploma thesis³⁷, I have implemented DASMI support for Dazzle, which will briefly be described in the following section. The DASMI support for ProServer and MyDAS has been implemented by Robert Finn and the DAS community, respectively, in an effort to make all major DAS servers DAS1.53E capable¹⁸¹.

Dazzle DAS server All modern DAS servers are implemented in a modular fashion, which will be exemplified in the following by describing our extension to Dazzle. The Dazzle server has originally been implemented by Thomas Down at the Wellcome Trust Sanger Institute and is now part of the BioJava project³⁰². When our work on DASMI started, Dazzle was the only available Java-based DAS server. Dazzle works as a Java Servlet and is composed of classes that can be categorized as core, handler, and data source classes. The core classes provide essential DAS functionality, such as initializing server components or generally handling DAS requests and responses. Handler classes are specific to the different DAS commands and are responsible for recognizing the particular query parameters and generating the associated XML response formats. Data source classes provide access to the actual data that are to

be made available as a DAS resource, for example, by connecting to a database or reflecting the specifics of a certain data format. The configuration of a particular server instance via appropriate handlers and data sources is done via an XML configuration file.

Implementation details of the Dazzle DASMI extension In order to add DASMI support to Dazzle, I implemented the new handler class named `InteractionHandler` and a number of data source classes for providing access to interaction data residing in different file formats like HUPO-PSI-MI XML2.5 (see [Section 2.5.2](#)), the Simple Interaction Format (SIF) used by Cytoscape³⁵², or in-house MySQL databases schema³⁷. To simplify working with different interaction data formats, I developed a common data model that is used to represent interaction data within Dazzle. This data model is tightly linked to the DASINT XML Schema and is thus composed of object classes named `Interaction`, `Interactor`, `Participant`, `Detail`, `Range`, and `Sequence`, with the object variables representing the XML attributes and values.

The interplay of all Dazzle components is best illustrated by a schematic example. When an `interaction` request is received by a DASMI server instance, the core classes forward this request to the `InteractionHandler`. The handler extracts the `interactor` and `detail` query parameters and issues a request to the particular data source class that has been associated with the requested DAS source in the Dazzle configuration file. The data source class retrieves the requested interaction data and returns them to the handler as a collection of `Interaction` objects. In the ultimate step, the `InteractionHandler` transforms this collection into a DASINT XML document, which is wrapped inside a DAS header and returned to the requester.

Most of the interaction data that have been collected in different projects within the MPIO^{313,343} are stored in MySQL database tables. The data source classes I implemented to make these data available as DASMI sources, `MPIPPReferenceSource` and `MPIDDIReferenceSource`, make use of a technique named database connection pooling to efficiently retrieve binary protein-protein interactions (PPIs) and domain-domain interactions (DDIs), respectively. In a database connection pool, a number of connections to the database are constantly maintained in an open and active state and can be requested from the application. This is in contrast to the classic approach of opening a database connection prior to a request and closing it right after. Especially in situations, where a large number of databases requests are made within a short period of time, the overhead produced by opening and closing those connections can lead to performance bottlenecks.

For the data source classes that retrieve interaction data directly from files like PSI-MI XML 2.5 documents, I implemented a caching solution based on (embedded) database support, ensuring that the performance of a DASMI server is sufficiently fast even when a large number of requests are made within a short time. Assuming that data are stored on normal hard-drives and not entirely kept in the main memory, the access on normal files is considerably slower than data retrieval from

indexed databases. Therefore, I developed two versions of database caching, one enabling access to user-uploaded PSI-MI XML documents in the DASMIweb client (see [Section 3.2.3.1](#)), and the `PSIMIReferenceSource`, which is distributed with `Dazzle` and intended for users who want to set up their own DASMI server instance. The difference between the two versions is that the latter uses an embedded database, which is initialized and filled at server startup, while the DASMIweb version uses standard in-house MySQL databases of the MPII. The embedded database ensures that `Dazzle` can be used even if there is no complex database management system installed on a server.

3.2.2.2 Data sources available via DASMI

Tables [3.3](#) list the sources that are available via DASMI servers as of November 2013. We developed DASMI to support different levels of molecular interactions and demonstrated this by providing a range of protein-protein and domain-domain interaction datasets, which have been determined with experimental and computational approaches on a large- and small-scale. The majority of DASMI servers is hosted at the MPII. Many of the sources contain final datasets that are not changed or updated, while some servers provide snapshots of data that are constantly updated, for example, IntAct or MINT. Ideally, these DASMI sources would not be hosted at the MPII but directly by the responsible institutions like (see [Section 3.4](#)).

3.2.3 DASMI client

A DASMI client is a gateway to distributed interaction data, allowing users to seamlessly communicate with multiple DASMI sources, without requiring any knowledge about the underlying data exchange protocol. In general, after a user requests a particular query, a DASMI client will communicate with the suitable DASMI sources, retrieve and unify their results, and present them to the user. During my Diploma thesis, I developed the prototype of a web-based client capable of unifying a range of molecular interaction types from DASMI servers³⁷. Subsequently, additional features have been added to this client, eventually resulting in a dedicated publication⁴¹. In the following sections, I will describe this tool named DASMIweb^e in more detail.

3.2.3.1 DASMIweb

DASMIweb was developed to act as a starting point for interactome studies, by consolidating molecular interactions from various resources. Our primary design goals, when developing DASMIweb, were interactivity and general usability. The former means that the user should have instant feedback and should not experience unnecessary delays, for example, allowing to change the DASMI server composition by

^e<http://www.dasmi.de/web/>

Table 3.3: DASMI sources providing domain-domain interactions (DDIs) and protein-protein interactions (PPIs). DDI sources all have the coordinate system Pfam,Protein Sequence. PPI sources can either be queried by UniProt accession numbers or Entrez Gene identifiers.

Data source name(s)	Reference	Coordinate system
3did	357	Pfam,Protein Sequence
APMM1, APMM2	401	Pfam,Protein Sequence
DIMA-DPEA, DIMA-DPROF, DIMA-STRING	282	Pfam,Protein Sequence
DPEA	325	Pfam,Protein Sequence
InterDom	261	Pfam,Protein Sequence
iPfam	116	Pfam,Protein Sequence
IPPRI, IPPRI-hc	339	Pfam,Protein Sequence
LDSC, LDSC-core	213	Pfam,Protein Sequence
LLZ	224	Pfam,Protein Sequence
LP	143	Pfam,Protein Sequence
PINS	43	Pfam,Protein Sequence
RCDP50	186	Pfam,Protein Sequence
RDFE	68	Pfam,Protein Sequence
Wuchty	409	Pfam,Protein Sequence
Bioverse, Bioverse-core	239	Entrez, Gene_ID
CCSB-HI1	327	Entrez, Gene_ID
DIP	333	UniProt, Protein Sequence
HiMAP, HiMAP-core	322	Entrez, Gene_ID
HomoMINT	290	UniProt, Protein Sequence
HPRD	193	UniProt, Protein Sequence
MDC	360	Entrez, Gene_ID
MINT	65	UniProt, Protein Sequence
OPHID	50	UniProt, Protein Sequence
POINT	166	Entrez, Gene_ID
Sanger, Sanger-core	215	Entrez, Gene_ID

adding new resources and instantly seeing the changes in the web interface (see [Figure 3.2.3.1.2](#)). Usability comprises features like automatic input identifier detection (see [Section 3.2.3.1.1](#)) or the clean and intuitive, yet configurable, user interface.

Figure 3.4: Graphical user interface of the web-based client DASMIweb. The client is embedded in the MPII website, centered between the header (top), footer (bottom) and the general navigation (left). The welcome view shown here is split into a *Query Panel* on the left and an *Information Panel* to the right. The blue box on the right can be used to show the *myDASMI Panel* (see [Figure 3.6](#)).

DASMIweb has been implemented as a three-layer application, consisting of a data access, business, and presentation layer. The presentation layer describes the graphical user interface (GUI), the only part of the application directly exposed to the user (see [Figure 3.4](#)). The programming logic, which defines how data are retrieved, unified, or temporarily stored, is contained in the business layer. The physical data access is defined in the data access layer. In the following, I will concentrate on selected features of the presentation and the business layers, as data access via our DAS protocol extension has already been detailed in [Section 3.2.1](#).

For the communication between presentation and business layer, we build upon a technology named Asynchronous JavaScript and XML (AJAX). Compared to a classic web-application, where a request results in a reload of the complete website, AJAX allows for replacing or updating parts of a website without refreshing the whole page. Today, such dynamic web-applications are ubiquitous, a prominent example being the panning view within Google Maps^f, which dynamically loads the map tiles that are currently visible on a user screen. In DASMIweb, an example for the use of AJAX is the gradual generation of the interaction table as soon as the

^f<http://maps.google.com/>

DASMI servers respond, which allows the compensate for different server response times. The AJAX functionality in DASMIweb is provided by a software library named Direct Web Remoting (DWR^g).

3.2.3.1.1 Identifier mapping Proteomics affords a great diversity of identifiers to describe genes, proteins, or other biological entities^{76,424}. As a result, interaction databases use a variety of identifier systems (the equivalent of coordinate systems in the DAS community) for describing their data³¹³. In order to request and unify results from different servers, DASMIweb converts between different identifiers system by using internal mapping tables derived from iProClass¹⁶⁴ and Pfam¹¹⁸. DASMIweb supports the following identifier systems for protein interactions:

- Entrez Gene identifier (e.g. 3064)
- UniProt accession number (e.g. P42858)
- UniProt identifier (e.g. HD_HUMAN)
- GeneInfo identifier (e.g. 30582895)
- RefSeq identifier (e.g. NP_002102.4)
- Ensembl Gene identifier (e.g. ENSG00000100296)

These identifier systems are compatible in the sense that mappings between them exist⁷⁶. For example, if a user requests interactions for the Huntingtin protein by querying with the UniProt identifier "HD_HUMAN", DASMIweb will convert this identifier to the UniProt accession number "P42858", the Entrez Gene identifier "3064" and other compatible identifier systems, in order to query interaction data from DASMI servers in their supported identifier or coordinate systems. After the individual DASMI servers have reported their results, a second mapping step is performed to unify and present the data to the user. In this mapping step, all gene and protein identifier systems are converted to Entrez Gene IDs. The reason for this mapping is that there is usually a one-to-one mapping from UniProt or Ensembl Gene ID to Entrez Gene ID, whereas there are multiple mappings for the opposite direction.

Fortunately, domain-domain interactions (DDIs) are predominantly reported using stable Pfam identifiers. Therefore, no identifier mapping is needed for DDI queries.

Querying One of our usability-related goals was to simplify querying DASMIweb by only using a single search box. Instead of asking the user to explicitly specify the identifier system of the input, DASMIweb tries to determine it automatically and

^g<http://getahead.org/dwr/>

DASMIweb - dynamic online integration and annotation of molecular interaction data

Query

1213

e.g. Entrez Gene, GI, Pfam, RefSeq, UniProtKB

Query Information

The query 1213 describes different biological entities.

We found 4 entities in our database that contain 1213 as their identifier or as part of their name. Please select the correct candidate below.

Potential interactors				
	ID	Description	Database links	Other details
<input type="button" value="Query"/>	1213 (Entrez Gene)	Clathrin heavy chain 1 (CLH-17).	UniGene: Hs.491351, Hs.663896. UniProtKB: CLH1_HUMAN, Q49ALO_HUMAN. Entrez Gene: 1213. UniParc: UPI0000127ABD, UPI000056F16E. IPI: IPI0002406Z, IPI00455383, RefSeq: NP_004850.1, NM_004859, Entrez GeneInfo: 32451593, 4758012, 30353925, 40788952, 119614805, 29983, 34364629, 119614802, 1705916, 71297093, UniProtKB: Q00610, Q49ALO, Ensembl: ENSG00000141367,	species: 9606
<input type="button" value="Query"/>	1213 (Entrez GeneInfo)	Hemoglobin subunit beta-C (Hemoglobin beta-C chain) (Beta-C-globin).	UniGene: Oar.14582, UniProtKB: HBBC_SHEEP, Entrez Gene: 100134870, UniParc: UPI000016C4AF, RefSeq: NP_001106896.1, NM_001113425, Entrez GeneInfo: 1213, 164663752, 62901571, UniProtKB: P68056,	species: 9940
<input type="button" value="Query"/>	258900 (Entrez Gene)	Olfactory receptor 1213 .	UniGene: Mm.389283, UniProtKB: A2ATG2_MOUSE, Q7TR07_MOUSE, Q8VGF8_MOUSE, Entrez Gene: 258900, UniParc: UPI00001951A0, UPI00001AA680, UPI00000291FD, RefSeq: NP_667109.1, NM_146898, IPI: IPI00466996, Entrez GeneInfo: 123233312, 81912238, 32073769, 81915784, 18479678, 22129091, UniProtKB: A2ATG2, Q7TR07, Q8VGF8, Ensembl: ENSMUSG00000075111,	species: 10090
<input type="button" value="Query"/>	Q7P924 (UniProtKB)	Hypothetical protein rsib_orf.1213 .	UniProtKB: Q7P924_RICSI, UniParc: UPI0000189405, RefSeq: ZP_00142960.1, Entrez GeneInfo: 34581480, 28262865, UniProtKB: Q7P924,	species: 272951

Figure 3.5: Querying and identifier mapping in DASMIweb. The user is asked to resolve the ambiguous query input "1213", which can either represent an Entrez Gene or GeneInfo identifier, but is also part of two additional gene descriptions.

only asks for confirmation in case of ambiguities (see Figure 3.5). DASMIweb can be configured to use partial or exact query string matching. In the default, partial matching, an exemplary query for "1213" would result in four potential interactors, as this query string is a valid Entrez Gene and GeneInfo identifier, but is also part of two gene descriptions. The latter two options would not be shown if the search behavior is changed to exact query string matching.

3.2.3.1.2 Interaction view The main view of DASMIweb presents all interaction results for a query interactor (see Figure 3.6). Like all views in DASMIweb, the interaction view comprises the *Query* and *Information Panel*. However, the latter now contains a brief description of the query interactor, listing database identifiers, names, and synonyms. The interactions are presented in a table in the central *Interaction Panel* below.

Visualization Each interaction partner of the query interactor, or multiple partners in the case of protein complexes, is listed in a separate row. The first three

Query

P51587

e.g. [Ensembl](#), [Entrez Gene](#), [GeneInfo](#), [Pfam](#), [RefSeq](#), [UniProtKB](#)

Interactor Information

Identifier: [P51587 \(UniProtKB\)](#) (Breast cancer type 2 susceptibility protein (Fanconi anemia group D1protein).) (breast cancer 2, early onset)

[UniProtKB](#) [BRCA2_HUMAN](#)

[UniGene](#) [Hs.34012](#)

[Entrez Gene](#) [675](#)

[UniParc](#) [UPI0000053473](#)

Database links: [Entrez](#) [119395734](#), [28400649](#), [1177438](#), [14424438](#), [23](#)

Quick Configuration

- Show Source Configuration
- Show Empty Sources: YES
- Set tabbing width: 10
- Clear Interactions

144 interactions (showing 1 to 10)

Name	ID	Description	SANGER-CORE	INTRACT	CCSB-H11	HIPRO	BIOVERSE	SANGER	MINT	HOMOINT	HITAP	HITAP-CORE	POINT	OPHID	MOC	DIP	BIOVERSE-CORE
FACD	2177	Fanconi anemia group D2 protein (Protein FACD2).															
	EBI-539895																
Q9BXW9-2	Q9BXW9-2																
p59-Fyn	2534	Proto-oncogene tyrosine-protein kinase Fyn (EC 2.7.10.2) (p59-Fyn) (Protooncogene Syn) (SLK).															
RAD51A	5888	RAD51 homolog (RecA homolog, E. coli) (<i>S. cerevisiae</i>)															
DSS1	7979	26 proteasome complex subunit DSS1 (Split hand/foot malformation type1 protein) (Deleted in split hand/split foot protein 1) (Splithand/foot deleted protein 1).															
Replication factor A protein 1	6117	replication protein A1, 70kDa															
RPA34	6118	replication protein A2, 32kDa															
ASH	2885	Growth factor receptor-bound protein 2 (Adapter protein GRB2) (SH2/SH3adapter GRB2) (Protein Ash).															
ABL	25	Proto-oncogene tyrosine-protein kinase ABL1 (EC 2.7.10.2) (p150) (c-ABL) (Abelson murine leukemia viral oncogene homolog 1).															
STAT5A	6776	Signal transducer and activator of transcription 5A.															

Figure 3.6: DASMIweb interaction view. The top right *Information Panel* describes the query interactor, the central *Interaction Panel* below contains all interactions in a tabular representation. Table columns identify DASMI sources that have been queried; colors indicate if the source provides experimentally determined data (green) or computational predictions (yellow). Table rows identify interaction partners; particular interactions are represented by blue squares at the intersections of rows and columns. The small *myDASMI Panel* on the right border allows setting visualization details like the number of interactions that are shown per page.

columns contain the Entrez Gene or Pfam identifier(s), name(s), and description(s). A separate column represents each DASMI source that has been queried for interactions thereafter. The source label in the table header lists the number of reported interactions and color-codes the type of interactions the data source provides: a green background indicates data that have been experimentally determined or curated from the scientific literature, a yellow background represents datasets that have been computationally predicted. Individual interactions between the query interactor and the interaction partner(s) in a row are marked by a blue square at the intersection of the row and the column of the DASMI source reporting the interaction (see [Figure 3.6](#)).

Customization The interaction table is built gradually. As soon as a server reports its results, new interactions are added to the end of the table. The table may be

sorted according to individual columns or a consensus over all DASMI sources. The latter allows for a quick visual inspection of the interaction results, based on the assumption that interactions that are reported by different databases have a higher likelihood of being true positive interactions. For the sake of clarity, sources that did not return any interaction can be excluded from the results table. A tabbed display allows to select the number of interactions that are shown per page, by default set to 50. All display and sorting options can be adjusted in the *myDASMI Panel*, which can be displayed by clicking on the small label on the right screen border (see Figures 3.4 and 3.6).

Interaction details Additional information about a particular interaction and the interaction partners involved may be requested by clicking on any of the interaction squares (see Figure 3.7). The data presented this way may contain links to the original publication or the full interaction record in the source database, information about the binding interfaces, or details on the experimental conditions. All details are provided by the DASMI sources, thus the amount of information may vary.

Data export All interaction results may be exported from DASMIweb in order to make them available for further analyses. Currently available data export formats are MITAB2.5 (see Section 2.5.3) and the Simple Interaction Format (SIF), a flat file format that is in particular supported by the network analysis and visualization platform Cytoscape³⁵².

Source configuration DASMIweb retrieves a list of available DASMI sources from the DAS registry. This list is combined with an internal list in order to incorporate sources that are not officially registered or that are web-services not following the DASMI protocol (see Section 3.3). Without any further configuration, DASMIweb will include all compatible interaction data sources into a user query. The user can change this selection in the *Source Configuration Panel*, which can be requested via a link in the *Query Panel* (see Figure 3.8).

Sources are grouped based on their coordinate or identifier systems. For each source, basic information like the name, URL, or a brief categorization of its data is provided. The latter is also indicated by the background color. As in the aforementioned source label in the interaction table (see Figure 3.6), green indicates data that have been experimentally determined, yellow represents computational predictions. Blue is used for servers that provide interaction confidence scores (see Section 4.2.1). By default, all compatible sources are active, indicated by a checkmark in the respective entry. By removing this checkmark, the source will be deactivated, indicated by the lighter background color, and will not be used for further queries. If a source is not responding to a query or is responding with errors, it will also be deactivated.

Query: P51587 Query

e.g. Ensembl, Entrez Gene, GeneInfo, Pfam, RefSeq, UniProtKB

Show Source Configuration

Interactor Information

Identifier: P51587 (UniProtKB) (Breast cancer type 2 susceptibility protein (Fanconi anemia group D1protein).) (breast cancer 2, early onset)

UniProtKB BRCA2_HUMAN,
UniGene Hs_34012,
Entrez Gene 675,
UniParc UPI0000053473,
Database links: Entrez 119395734, 28400649, 1177438, 14424438, 2315186, 27065822

144 interactions (showing 1 to 50) Export as ... Select confidence measure ... Query Confidence Sources

Name	ID	Description	SMER-CORE	INTACT	CCSB-H11	HPRD	BIOVERSE	SMER	MINT	HOMOINT	HIMAP	HIMAP-CORE	POINT	OPRID	HDC	DTP	BIOVERSE-CORE
FACD	2177	Fanconi anemia group D2 protein (Protein FACD2).															
	EBI-539895																
Q9BXW9-2	Q9BXW9-2																
p59-Fyn	2534	Proto-oncogene tyrosine-protein kinase Fyn (EC 2.7.10.2) (p59-Fyn) (Protooncogene Syn) (SLK).															
RAD51A	5888	DNA repair protein RAD51 homolog 1 (hRAD51) (HsRAD51).															

Details from source IntAct

link <http://www.ebi.ac.uk/intact/binary-search/faces/search.xhtml?query=EBI-297231>

interaction type interaction type

detection method x-ray diffraction

authors Pellegrini et al. (2002)

PubMed 12442171

original interaction name P51587-Q06609

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/intact/interaction?interactor=P51587>

Details from source HPRD

link http://www.hprd.org/interactions?protein=01557&isoform_id=01557_1&isoform_name=

link http://www.hprd.org/interactions?protein=02554&isoform_id=02554_1&isoform_name=

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/hprd/interaction?interactor=P51587>

Details from source HomoMINT

link <http://mint.bio.uniroma2.it/HomoMINT/search/search.do?queryType=protein&ac=675>

original interaction name 675-5888

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/homomint/interaction?interactor=675>

Details from source HiMAP

original interaction name 675-5888

link <http://www.nature.com/nbt/journal/v23/n8/sbs/nbt1103.html>

originalConfidenceScore 761

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/himap/interaction?interactor=675>

Details from source POINT

original interaction name 675-5888

link <http://bioinformatics.oxfordjournals.org/doi/content/abstract/20/17/3273>

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/point/interaction?interactor=675>

Details from source MINT

link <http://mint.bio.uniroma2.it/mint/search/search.do?queryType=protein&ac=P51587>

DAS query string <http://dasmi.bioinf.mpi-inf.mpg.de/das/mint/interaction?interactor=P51587>

Additional details for interactor Q06609:

species 9606

Description DNA repair protein RAD51 homolog 1 (hRAD51) (HsRAD51).

type protein-coding

description RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)

synonyms BRCC5|HRAD51|HsRad51|HsT16930|RAD51A|RECA

chromosome 15

location 15q15.1

db_xref:uniprot_id RAD51_HUMAN, B0FXP0_HUMAN, B2R8T6_HUMAN, B4DZT8_HUMAN, Q5U0A5_HUMAN, Q6TAR4_HUMAN, Q6ZNA8_HUMAN, Q9NZG9_HUMAN

db_xref:unigene Hs_631709

db_xref:entrez 5888

db_xref:uniparc UPI000000D8D2, UPI0000EE6B91, UPI00017A8568, UPI0000453989, UPI0000212EE0, UPI000035E763, UPI000006D4C2, 5733658, 154275771, 548663, 27065821, 19924133, 12655203, 49168602, 397827, 27368250, 6730074, 7767554, 285977, 164506989, 119612838, 119612840, 189069251, 194381662, 54696278, 38017105, 158257370, 47077076, 119612841, 7407071

db_xref:refseq NP_002866.2, NP_597994.2, NM_002875, NM_133487

db_xref:uniprot_ac Q06609, B0FXP0, B2R8T6, B4DZT8, Q5U0A5, Q6TAR4, Q6ZNA8, Q9NZG9

db_xref:ipii IPI00032201, IPI00220649, IPI00553199

db_xref:Ensembl ENSG00000051180

DSS1 7979 26 proteasome complex subunit DSS1 (Split hand/foot malformation type1 protein) (Deleted in split hand/split foot protein 1) (Splithand/foot deleted protein 1).

Figure 3.7: Part of the DASMIweb interaction view showing additional details for the interaction between the query gene BRCA2 with RAD51A, as reported by IntAct. Interaction details from other sources, which also report this interaction, are listed in grey, followed by details on the interaction partner RAD51A.

Interaction source selection					
<input type="button" value="Add New Source"/> <input type="button" value="Entrez, Gene_ID"/> <input type="button" value="UniProt, Protein Sequence"/> <input type="button" value="Pfam, Protein Sequence"/>					
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Name	Description	URL	Labels
<input checked="" type="checkbox"/>		IntAct	Kerrien et al. (2007). IntAct-open source resource for molecular interaction data. <i>Nucleic Acids Res</i> , 35(Database issue):D561-D565.	http://www.ebi.ac.uk/intact/psicquic/webservices/psicquic/	<ul style="list-style-type: none"> ■ literature curated ■ psicquic
<input type="checkbox"/>		MPIDB	Goll et al. (2008). MPIDB: the microbial protein interaction database. <i>Bioinformatics</i> , 24(15):1743-1744;	http://www.jcvi.org/mpidb/servlet/webservices/psicquic/	<ul style="list-style-type: none"> ■ literature curated ■ psicquic
<input checked="" type="checkbox"/>		HomoMINT	Persico et al. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. <i>BMC Bioinformatics</i> , 6 Suppl 4:S21.	http://dasmi.bioinf.mpi-inf.mpg.de/das/homomint/	<ul style="list-style-type: none"> ■ predicted ■ temporarily cached ■ version: 2009-03-05
<input checked="" type="checkbox"/>		HPRD	Keshava Prasad et al. (2009). Human Protein Reference Database-2009 update. <i>Nucleic Acids Res</i> , 37(Database issue):D767-D772	http://dasmi.bioinf.mpi-inf.mpg.de/das/hprd/	<ul style="list-style-type: none"> ■ literature curated ■ temporarily cached ■ 2009-06-07
<input checked="" type="checkbox"/>		DIP	Salwinski et al. (2004). The Database of Interacting Proteins: 2004 update. <i>Nucleic Acids Res</i> , 32:D449-D451	http://dasmi.bioinf.mpi-inf.mpg.de/das/dip/	<ul style="list-style-type: none"> ■ literature curated ■ temporarily cached ■ version: 2009-01-26
<input type="checkbox"/>		OPHID	Brown and Jurisica (2005). Online predicted human interaction database. <i>Bioinformatics</i> , 21(9):2076-2082.	http://dasmi.bioinf.mpi-inf.mpg.de/das/ophid/	<ul style="list-style-type: none"> ■ predicted ■ temporarily cached ■ version: 2009-08-03
<input checked="" type="checkbox"/>		FunSimMat2.1	Schlicker and Albrecht. (2008). FunSimMat: a comprehensive functional similarity database. <i>Nucleic Acids Res</i> , 36, D434-D439. @BPscore: The BPscore is based on biological process annotation of the Gene Ontology. Range: 0-1 Link: http://funsimmat.bioinf.mpi-inf.mpg.de/help.php @CCscore: The CC score is based on the cellular component annotation of the Gene Ontology. Range: 0-1 Link: http://funsimmat.bioinf.mpi-inf.mpg.de/help.php @MFScore: The MF score is based on the molecular function annotation of the Gene Ontology. Range: 0-1 Link: http://funsimmat.bioinf.mpi-inf.mpg.de/help.php	http://dasmi.bioinf.mpi-inf.mpg.de/das/funsimmat/	<ul style="list-style-type: none"> ■ interaction quality measure
<input checked="" type="checkbox"/>		Domain support	@predicted: This measure indicates protein-protein interactions that are supported by underlying domain-domain interactions. The domain-domain interaction datasets used in the predicted subset are based on diverse computational predication algorithms. Please see the respective interaction details for more information on the datasets. Range: 1 Link: http://www.dasmi.de @crystal-structure: This measure indicates protein-protein interactions that are supported by underlying domain-domain interactions. The domain-domain interaction datasets used in this subset are based on the analyses of PDB crystal-structures. Please see the respective interaction details for more information on the dataset. Range: 1 Link: http://www.dasmi.de	http://dasmi.bioinf.mpi-inf.mpg.de/das/domains/	<ul style="list-style-type: none"> ■ interaction quality measure

Remove All Sources Load Sources From Registry Close Configuration

Figure 3.8: DASMIweb *Source Configuration Panel*. Interaction sources are retrieved from the DAS registry and an internally maintained list. Sources are sorted according to their coordinate systems and are described by basic information like name and URL. Sources with a leading checkmark are active, inactive sources do not have the checkmark and are indicated by a lighter background color. Green backgrounds indicate sources that provide experimentally determined or manually curated data, yellow represents computational predictions, blue is used for sources that provide interaction confidence scores.

Including new data DASMIweb supports different means for adding new interaction sources. If a DASMI source is registered at the DAS registry, it will automatically be available to all DASMIweb users without any further efforts. Another option, for instance, if the source and its data should not be released to the general public, is the local registration of an existing DASMI source via the respective section of the DASMIweb *Source Configuration Panel*. These two approaches require that the data that are to be included have already been made available as a server. The difference is that the registration at the DAS registry affects all DASMIweb users, while a local registration will only be available to the single user. A third option is to upload a PSI-MI XML2.5 file and make its content available in the user session of DASMIweb. This last option allows for comparing unpublished data with the results from public data sources or to annotate these data with interaction confidence scores.

3.2.3.1.3 Interaction confidence scoring A feature that is only briefly mentioned here, as it is described in more detail in [Chapter 4](#), is the ability of DASMIweb to assess interaction data quality via different measures. In addition to confidence scores already provided by the source interaction datasets, DASMIweb can incorporate dedicated scoring servers, which are marked blue in [Figure 3.8](#). The retrieval of interaction confidence scores has to be explicitly requested by the user in order to minimize unnecessary computational overhead. This can be done by clicking on the corresponding button at the top of the *Interaction Panel*. DASMIweb lists all scores in the detail section of a particular interaction (see [Figure 3.7](#)) and additionally uses a graphical representation via different node colorings to afford an intuitive assessment (see [Figure 4.2](#)).

3.2.3.2 *iPfam graphical domain interaction browser*

The iPfam graphical domain interaction browser^h is a DASMI client that has been developed by the iPfam group of Robert Finn at the Wellcome Trust Sanger Institute. iPfam is a database of Pfam domain interactions derived from proteins with an experimentally determined 3D structure¹¹⁶. The graphical domain interaction browser uses DASMI to complement the results of the iPfam database with those from other domain-domain interaction resources (see [Table 3.3](#)). For a user-selected Pfam domain, the iPfam browser retrieves results about interacting domains from a number of user-selected DASMI servers and represents them in a graphical network view (see [Figure 3.9](#)). By clicking on a node in this network, the interaction results for this domain are requested, allowing a quick exploration of the extended domain interaction network. Unfortunately, since Robert Finn moved to the Howard Hughes Medical Institute (HHMI) Janelia Farm in 2010, taking the iPfam project with him, the URL of the graphical domain browser is no longer active and the

^h<http://ipfam.sanger.ac.uk/graph>

resource currently not available.

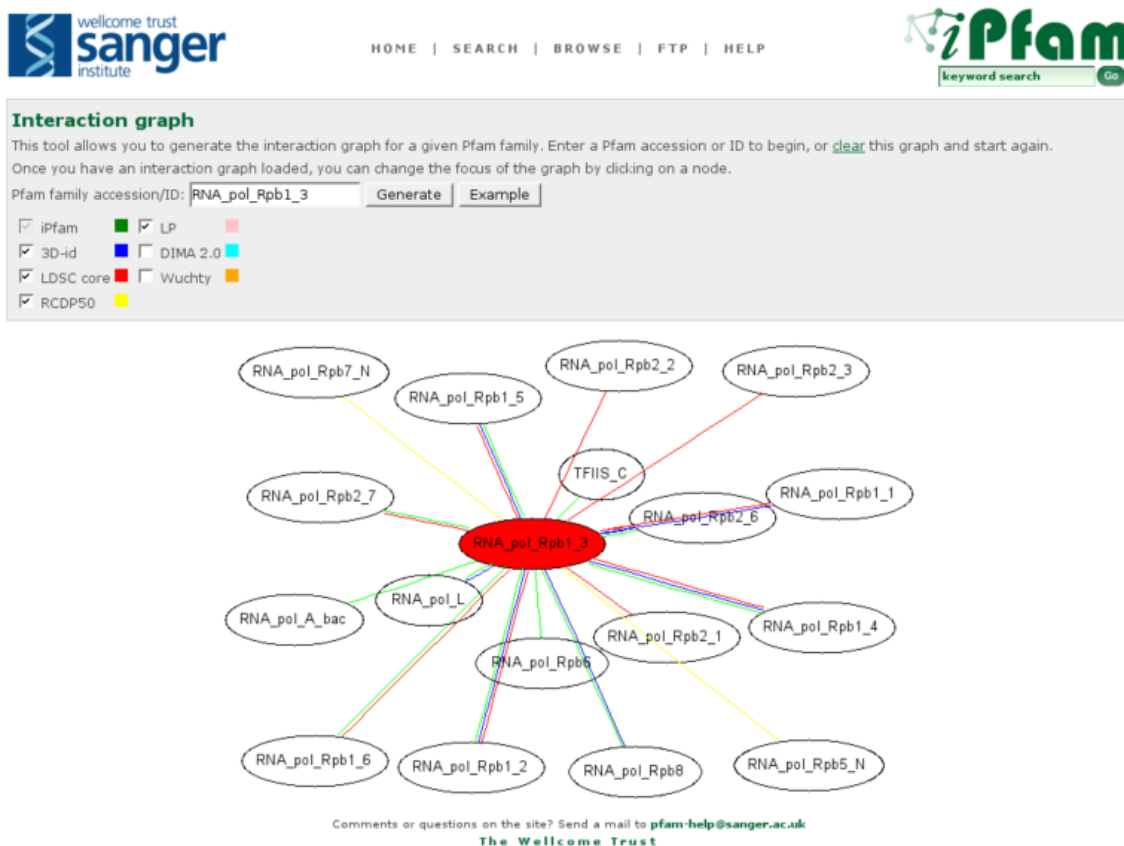


Figure 3.9: The iPfam graphical domain interaction browser combines various domain-domain interaction sets in a graphical network view⁴⁰. Protein domains are depicted by ovals, interactions as connecting edges. Edge colors distinguish individual data sources. By clicking on a node, the interactions for this particular domain are requested.

3.3 PSI Common Query Interface

The Proteomics Standards Initiative Common Query Interface (PSICQUIC) is a decentralization approach for molecular interaction data retrieval that is very similar to DASMI¹⁴. While the idea for this distributed system dates back to the time when we started working on DASMI (see Section 3.2), its realization was delayed and it was still merely a concept when DASMI was officially published^{40,191,268,276}. The PSICQUIC architecture is very similar to DASMI. It too consists of a data exchange specification, servers providing interaction data, clients for data retrieval and unification, and a registry that maintains a list of available servers.

The major difference between DASMI and PSICQUIC is the data exchange specification. While DASMI has its origin in the DAS community, PSICQUIC was started

as a HUPO-PSI project to standardize access to interaction data. It is thus built around the PSI-MI formats²⁷².

PSICQUIC web service Web services such as PSICQUIC are traditionally accessed using the Simple Object Access Protocol (SOAP), an XML-based official W3C standard that may be used to transmit arbitrary messages over a transport protocol like HTTP. The interface of a SOAP web service, that is, the methods it provides and the data and parameters it accepts are defined using another XML dialect named Web Service Description Language (WSDL). In addition to XML-based requests via SOAP, PSICQUIC also supports queries via specific URLs. These URLs are based on Representational State Transfer (REST), a paradigm that defines a certain server output in response to a particular URL request. PSICQUIC REST queries are comparable to DASMI interaction requests (see [Section 3.2.1.1](#)).

The initial PSICQUIC specification¹⁴ⁱ, defined the following methods for data retrieval:

- `getByInteractor` and `getByInteractorList` retrieve interactions via database identifiers of the query interactor(s). These queries are thus similar to the `interaction` request in DASMI (see [Section 3.2.1.2](#)).
- `getByInteraction` and `getByInteractionList` allow for retrieving particular interactions by their identifier(s).
- `getQuery` enables more complex queries via the Molecular Interaction Query Language (MIQL, see [Section 3.3](#)).
- `getSupportedDbAcs`, `getSupportedReturnTypes`, and `getVersion` can be used to retrieve meta information about the PSICQUIC service.

Molecular Interaction Query Language (MIQL) The Molecular Interaction Query Language (MIQL^{14j}) is a query language based on Apache Lucene^k. Compared to queries by interactor or interaction identifiers, MIQL is much more powerful, allowing complex queries using wildcards and logical operators. The exemplary MIQL query `"park2 AND cask AND "pull down"` would retrieve all interactions that involve the gene products of Parkin and Cask, which have been determined via pull down experiments. As this query uses gene names, it would not be restricted to human interactors, but would return interactions from other organisms.

MIQL queries can be restricted to certain fields of a MITAB file. For instance, the search string `"species:human"` could be added to the last example to filter unwanted species. In MIQL 2.6, which is based on MITAB 2.6, and later versions, the query `"park2 AND complex:spoke` would return interactions of Parkin, which have resulted

ⁱ<http://code.google.com/p/psicquic/>

^j<http://code.google.com/p/psicquic/wiki/MiqlReference>

^k<http://lucene.apache.org>

from the spoke expansion of a protein complex. In MIQL 2.5, the query would not be restricted, as the complex expansion method was not available in MITAB2.5 (see [Section 2.5.4](#)).

Reference implementation The HUPO-PSI, essentially led by the IntAct team of the European Bioinformatics Institute (EBI), has developed a reference implementation for PSICQUIC to facilitate the installation of new PSICQUIC servers¹⁴. This reference implementation is based on a MITAB file that contains all the interaction data. Via an indexing process, the content of this file is made available as a PSICQUIC server.

Initially, the reference implementation only supported MITAB 2.5. A new reference implementation, which was recently published⁸⁸, now also provides support for versions 2.6 and 2.7. However, still no support is available for PSI-MI XML files. In order to serve their content, these files still have to be transformed to MITAB documents, a transformation that is not fully reversible, as not all the information can be preserved. In particular, protein complexes cannot be described in MITAB without the prior application of a complex expansion method (see [Section 2.5.4](#)).

PSICQUIC has a strong support in the interaction community. The project started with 4 servers, when it was officially published in 2011, this number already grew to 16. As of December 2013, 28 servers are registered at the PSICQUIC registry, including most of the key databases in the field.

PSICQUIC registry The PSICQUIC registry, currently hosted at the EBI¹, maintains a list of all available PSICQUIC servers. Like in the DAS registry (see [3.1.1.4](#)), individual servers may be described with certain keywords, such as "spoke expansion", "predicted", or "imex curation" to allow for a general categorization of their data. Furthermore, the version of the reference implementation used and the number of interactions provided in total is listed. PSICQUIC clients may retrieve a list of servers that follow specific requirements programmatically.

PSICQUIC clients PSICQUIC clients allow access to interaction data that are provided via the PSICQUIC specification. If the REST interface of a single PSICQUIC server is used, MITAB data can directly be retrieved via a standard browser, not requiring a specific client. For individual computational access, the project website^m contains source code to facilitate the retrieval of interaction data via different programming languages. In addition, a number of clients exist, which can be used, without requiring any programming efforts, in particular, PSICQUIC view¹⁴, molecular interactions cluster (miclusterⁿ), the PSICQUIC Cytoscape³⁵² client^o, and a

¹<http://www.ebi.ac.uk/Tools/webservices/psicquic/registry/>

^m<http://code.google.com/p/psicquic/>

ⁿ<http://code.google.com/p/micluster/>

^o<http://apps.cytoscape.org/apps/psicquicuniversalclient>

BioConductor¹³⁴ package^P.

3.4 Conclusions

3.4.1 Summary

In this chapter, I presented two systems for the dynamic exchange and integration of molecular interaction data. The idea for these approaches was born, when the growing public attention for systems biology and the technological and methodological improvements in the area of interaction determination, have led to substantial amounts of interaction data being generated in numerous laboratories. While most of these data were made publicly available, they were not readily accessible to the research community, as they were scattered over a multitude of online resources. In contrast to the protein structure and sequence fields, where single large resources like PDB³⁴ or UniProt³⁷⁵ act as central data repositories, the collection of molecular interactions has traditionally been spread over a number of smaller resources¹⁸. These primary interaction databases, which acquire their data directly from experimentalists or by curating the scientific literature, are invaluable resources for proteomics. However, they are commonly focused on specific biological aspects like extracellular interactions or certain signaling pathways²⁷⁰.

In 2004, the HUPO-PSI introduced an XML format for the representation of interaction data, later complemented by a tabular representation^{153,277}. This fostered the development of integrative interaction databases, which collect and unify interaction data from a number of primary databases^{66,180,298,315}. These centralized repositories can be useful resources for researchers, as a great number of interaction data are available at a single site. However, in order to remain relevant and up to date, the unification has to be constantly repeated, as the source databases keep on updating and expanding. In addition, a central authority decides about the interaction data that are to be incorporated into these systems, for instance, limiting the focus on experimentally determined interactions and neglecting computational predictions.

The complementary approach to data centralization is federation, where the data are kept with their providers and are retrieved on request. This approach, successfully showcased by the Distributed Annotation System (DAS) for genome and proteome annotation, empowers the users, as they can decide, which data are to be incorporated or left out. In addition, data maintenance is greatly improved, as the original providers keep full control over their data.

In order to transfer the concept of data federation to the field of interaction data, we developed an extension to DAS named Distributed Annotation System for Molecular Interaction (DASMI^{37,40}). The basic idea of DASMI is to leave the interaction data with their original providers and collect them upon request. For the user, the system should act like a centralized repository, that is, the interaction data should

^P<http://www.bioconductor.org/packages/release/bioc/html/PSICQUIC.html>

be unified across the different data sources and be available at a single location. To this end, we developed DASMI as a client-server system, where interaction data are provided by multiple DASMI servers and are presented to the user in a DASMI client. The communication between client and server is handled by a data exchange protocol.

First, we required a data exchange protocol capable of describing different types of molecular interactions. We started by evaluating existing alternative data formats. The one with the largest community support, PSI-MI XML¹⁵³, can be used to describe molecular interactions with a fine level of detail, necessarily resulting in very complex, deeply branched documents. Since DAS was built around the idea of simplicity, and because tool support for processing PSI-MI documents was very basic at the time, we decided against the use of PSI-MI XML. Interestingly, even in 2013, PSI-MI XML is still not supported as input to the PSICQUIC reference implementation, although both are official standards of the HUPO-PSI⁸⁸. MITAB, on the other hand, which was originally published as a simple tabular alternative to the complex PSI-MI XML, became increasingly popular and has grown from 15 to 42 columns in its current version 2.7.

Our data exchange protocol, consisting of a new DAS command named *interaction* and the corresponding *DASINT* XML response format, was developed in the context of the European BioSapiens Network of Excellence³⁷⁷ and in close collaboration with the DAS community. Together with other DAS extensions, it was bundled into the unofficial DAS version 1.53E¹⁸¹. Following the definition of the data exchange specification, we developed a DAS server capable of providing interaction data in the new format. We started by extending the existing Java-based server library *Dazzle*^{94,302}, the DAS community later followed with extensions to the other major DAS servers *ProServer*¹¹⁷ and *myDAS*³²⁹. These servers can now be used to serve interaction data residing in different database schemas, flat files, and PSI-MI XML files. Using this extended *Dazzle* library we set up a number of DASMI sources providing experimentally determined or computationally predicted protein and domain interactions, which had been collected in previous projects at the MPII. All these servers were also registered at the DAS registry, making them available to every DAS user.

In order to retrieve, unify, and present the interaction data to the user, we developed the web-based client *DASMIweb*⁴¹. As the name suggests, this client is primarily based on the DASMI protocol, but can incorporate other web services like PSICQUIC. *DASMIweb* is a dynamic client that strongly depends on Asynchronous JavaScript and XML (AJAX) for ad-hoc data retrieval and presentation. In addition to binary protein-protein interactions and protein complexes, *DASMIweb* may also be used to incorporate knowledge about the interacting domains, thus helping to elucidate protein interaction networks.

DASMI was the first published approach for decentralized interaction data retrieval⁴⁰. It showcased the potential of data distribution for interactomics and has

successfully been used in a number of research projects^{60,92,199,206,295,387,407}. However, a major factor limiting a wider use of DASMI, in particular, a larger number of external data providers, was our data exchange format. While the DASINT format proved to be sufficient for its task and was well accepted by the DAS community¹⁸¹, it was not supported by the major interaction data providers, like those associated with the HUPO-PSI and the IMEx consortium²⁷⁸. These database providers instead developed their own distributed system for interaction data exchange, named PSI Common Query Interface (PSICQUIC²⁷⁶). The similarities of PSICQUIC and DASMI are immanent: both consist of interaction data servers, clients for presenting the results, a data exchange specification, and a registry keeping track of available resources¹⁴. As such, PSICQUIC can be seen as the more sophisticated successor to DASMI. PSICQUIC has a great support in the interaction community, as of December 2013, 28 servers are available at the PSICQUIC registry. In addition, PSICQUIC client libraries exist for various programming language, Cytoscape³⁵² and BioConductor¹³⁴ and the system is used for collaborative annotation efforts like IMEx^{14,271}.

3.4.2 Outlook

Further development on DASMI has largely stopped since PSICQUIC was published as its successor system. In particular, no additional work will be invested in DASMI servers or the data exchange protocol, despite potential improvements with respect to the supported input data formats. The client DASMIweb, however, may be extended to a general interactome browser. In particular, it already supports several features that no other PSICQUIC-enabled client provides, including the easy incorporation of unpublished, private interaction data or the visualization of interaction confidence scores. When DASMIweb was published in 2009, it already supported the then unpublished PSICQUIC standard⁴¹ but as no PSICQUIC registry was available at the time, selected servers had to be manually curated in an internal list. The first step to an improved PSICQUIC support would thus include the automatic access to the PSICQUIC registry. In addition, a closer integration of the Molecular Interaction (MI) controlled vocabulary would be beneficial for a better unification and presentation of the data.

PSICQUIC, as the official successor to DASMI, already has a strong support in the interaction community and is likely to continue its success story with more servers being made available. One area that still needs improvement is the unification of results from various servers. PSICQUIC view, for instance, only uses identifiers from UniProt, iRefIndex, DDBJ/ EMBL/ GenBank, RefSeq, and ChEBI for unification. Entrez Gene IDs, as they are used for gene-centric unification in DASMIweb, are not supported. Identifier mapping in general remains an important issue for PSICQUIC, for instance, limiting its query capabilities. The search index of a PSICQUIC server is currently built using only the information present in the input MITAB file. Some servers provide UniProt accession numbers, others use RefSeq or Entrez Gene

identifiers, with the result that their data can only be retrieved if a query of the corresponding identifier system is made. A query approach like implemented in DASMIweb, where a query identifier is first mapped to all compatible identifier systems, could prevent this.

Distributed systems like DASMI and PSICQUIC are generally not the first choice when genome-wide interaction data are required. Integrative data warehouses like iRefIndex, which provide unified interactome data, are largely used due to their faster analyses capabilities. However, with a suitable client, PSICQUIC could provide the basis for a data warehouse that can easily be maintained, a hybrid mixture between distributed and centralized systems⁵⁸. We are currently developing such an integrative interaction data warehouse, which is built upon individual MITAB files retrieved from PSICQUIC servers. Compared to ad-hoc PSICQUIC queries, this pre-built PSICQUIC-based warehouse allows for faster queries, an improved identifier mapping procedure, and above all better control of versioning and reproducibility.

A key requirement for PSICQUIC in the future is the support of PSI-MI XML files, such that protein complexes or cooperative interactions do not need to be expanded. Currently, these data have to be fitted in a tabular data format that was not designed for such complex interactions.

4 Assessing molecular interaction data quality

This chapter is dedicated to the assessment of interaction data quality, an area that, apart from general data availability, poses another challenge for interactomics. [Section 4.1](#) will describe the relevant data quality problems that are associated with different interaction determination and prediction techniques. This is followed by a brief description of the various confidence scoring approaches that have been proposed. In [Section 4.2](#), two systems will be presented that use the principle of data decentralization in order to make these different scoring algorithms easily available to the scientific community. Our first approach uses the Distributed Annotation System for Molecular Interaction, which was presented in [Section 3.2](#). Building on this proof-of-concept study, I will then describe a successor system that we designed in collaboration with the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI). I will end the chapter in [Section 4.3](#) by summarizing the current status and providing an outlook on potential future development

4.1 Introduction

[Chapter 3](#) described how distributed systems can be used to improve the availability of interaction data. A second major issue in interactomics, which despite intense research efforts is not adequately resolved, is data quality. While the error rate in genomic experiments is below the percentage level¹⁶⁹, large-scale protein interaction studies have been linked with false positive rates in the order of up to 50 percent^{86,148,252,396}.

4.1.1 Interaction data quality issues

Within the research community, individual techniques for determining molecular interactions are perceived differently with respect to their data quality and reliability. While some opinions and reservations are explainable, there is no profound scientific reason for others¹³³.

Yeast-two hybrid The experimental detection technique that has been particularly associated with varying data quality is the yeast-two hybrid system (Y2H, see [Section 2.2.1](#)). Y2H has been used in most of the early large-scale studies, and it was argued that the low overlap between similar studies performed in yeast^{171,382} or in human^{327,360} is an indicator of technical limitations^{163,165,208}. Self-activating reporter baits or the unphysiological overexpression of baits and preys in the yeast nucleus, are reasons that may lead to false positive results, spurious interactions that are wrongfully reported. Misfolded proteins or proteins requiring post-translational modifications that are not available in yeast, on the other hand, may result in false negative interactions that are not detected in a screen⁴²⁶. For reported interactions, it also remains questionable whether they are biological, that is, true positive interactions also present in the cell, or biophysical interactions that could take place, but may not³⁸⁸.

Affinity purification High-throughput studies using affinity purification followed by mass spectrometry (AP-MS, see [Section 2.2.1](#)), have long been considered to be superior to Y2H screens, in particular, because AP-MS screens may be performed under physiological conditions¹³⁰. Nevertheless, AP-MS also has its limitations, such as weak connections that might be diluted in the purification process and result in incomplete complexes³⁹⁶, or sticky proteins that might be purified regardless of the actual bait protein^{110,264}. In addition, the extracted purifications only indicate indirect co-complex relationships, which in the past have often been mistaken with true binary interactions like those detected with Y2H^{339,418}. Comparative studies have recently shown that the error rates of large-scale studies are much lower than initially anticipated and that no experimental protocol is superior to another^{49,388,421}.

Literature curation Manual curation of the scientific literature (see [Section 2.3.1](#)), in particular, curation of small-scale experiments that reported few interactions, has long been considered to be of higher quality than results obtained via large-scale experiments^{126,319,396}. Consequently, such data were used as gold standard in various studies^{123,214,309}.

In a much debated series of articles, literature curation has recently been associated with significant false positive rates, mostly caused by mistakes in human curation^{79,80,332}. According to [Venkatesan et al.](#), large-scale techniques like Y2H may even be of technically better quality than literature curation³⁸⁸. This statement is too general, however, as literature curation may be performed according to different quality standards³³². In addition to deep curation according to IMEx guidelines²⁷⁸, more shallow curation following the MIMiX guidelines²⁸¹ or curation following no particular recommendations^{79,193} is common. Unsurprisingly, the latter approach has been shown to have the highest error rates⁷⁹. Further complicating this distinction, IMEx databases like MINT also contain data that have not been curated according to the high IMEx standards²²⁰.

Compared to potentially less biased large-scale approaches³⁹⁶, literature curation is also faced with a study or selection bias, as there is significantly more literature on well studied or medically relevant genes, such as the famous cancer related BRCA2 and TP53, than there is for uncharacterized genes. This bias is then also reflected in the corresponding protein interaction databases^{80,125,252}.

Need for assessment Computational predictions (see [Section 2.2.2](#)) are generally approached with skepticism³⁹², despite their heterogeneous nature and reports that at certain specificity rates some approaches might yield data of a quality comparable to large-scale studies^{179,215}.

In summary, it can be seen that every interaction determination or prediction technique has certain strengths and weaknesses. Different approaches are required as they are complementing each other and are thus able to detect interactions that others are missing^{196,221,418}. Regardless of the technology that is used, however, errors are inevitable^{48,388}. It is thus evident that in order to achieve a high quality map of the human interactome, means for individual assessment of interaction data are crucial.

4.1.2 Methods for assessing molecular interactions

A great variety of methods have been developed for scoring individual protein interactions or complete datasets, for example, reviewed by Chua and Wong⁷². In the following, I will present a brief categorization of the techniques that are commonly employed. In general, the underlying methodology can be used for assessing interactions that have already been determined, but may also be used in order to predict interactions *de novo*³⁶⁷. Therefore, a significant overlap exists between the methods mentioned below and the computational prediction algorithms presented in [Section 2.2.2](#).

4.1.2.1 Experimental replication

It is common sense that an experimental finding that can be replicated in other experiments has a higher likelihood of being a true positive finding. In particular, it has been found that protein interactions that are reported in more than one study have a greater chance of being true positives^{215,355,396}. This approach has been employed by various experimental groups for creating *core* or *high confidence* subsets containing only interactions that appeared in multiple repetitions of an assay^{130,171,219,382}. Compared to such replications with the same assay, an even greater probability of being a true positive is assigned to interactions that can be replicated using a different experimental setup²¹⁵. This assumption also provides the basis for overlap computation, which is widely used for comparing interaction screens^{20,125,236,313,396}.

Experimental replication is incorporated into various scoring approaches. A popular example is the MINT-score²²⁰ that is computed based on all experimental ev-

idences and scientific publications that support a particular interaction. Weight factors are assigned to individual detection techniques, scoring direct interactions higher than co-complex associations, and the number of interactions a publication reports in total, granting higher weights to small-scale studies that reported at most 50 interactions than to large-scale studies²²⁰. An updated MINT-score is used in the integrative interaction database mentha⁵⁸ (see Figures 2.2 and 4.1). IntAct¹⁹⁰ also recently implemented a weighted replication based protocol to score experimentally determined interactions. In particular, this should filter out low-confidence interactions before exporting data to the UniProt consortium³⁷⁵ or the Gene Ontology Annotation project⁵⁹.

4.1.2.2 Gold standard

The term *gold standard*, sometimes also *golden standard*, is used for datasets that can be trusted to represent the biological truth, at least up to a reasonable extent¹⁷⁸. Transferred to interactomics, two gold standard or reference sets can be distinguished⁴⁸. First, the *positive gold standard* or *positive reference set* that describes true interactions. This set is commonly built as a high confidence subset of interactions that have been reported in multiple different experiments and that have been extensively validated^{49,86,215,367,388}. Second, the *negative gold standard* or *negative reference set* that should describe interactions that are not taking place *in vivo*³². The definition of this set is more difficult, as an interaction that is not reported in a database does not imply that the interaction is not taking place, it might just not have been detected. Approaches for defining negative reference sets are based on some forms of randomization^{86,388,426}, protein pairs with contradicting subcellular localizations^{52,179}, network distances²⁰, or experimental repetitions³⁷⁸. The group headed by Marc Vidal has recently proposed community efforts for refining common positive and negative reference sets that would allow different groups to globally calibrate and score their assays in order to make them more comparable^{49,388}.

4.1.2.3 Evolutionary conservation

The idea behind evolutionary conservation is that an interaction that is conserved across different species has a greater likelihood of being biologically relevant. Walhout et al. have introduced the concept of *interologs* for such orthologous proteins that have maintained their ability to interact in different organisms⁴⁰⁰. Evolutionary conservation has not only been used as a measure of data quality, but also to predict protein interactions in species without sufficient interaction data coverage, by transferring interologous interactions from model organism that are usually better studied^{166,215,399}. Lim et al. have found that, based on functional annotations, interologous predictions have a quality similar to data from literature curation or large-scale Y2H screens²²². Deane et al. proposed a similar concept based on par-

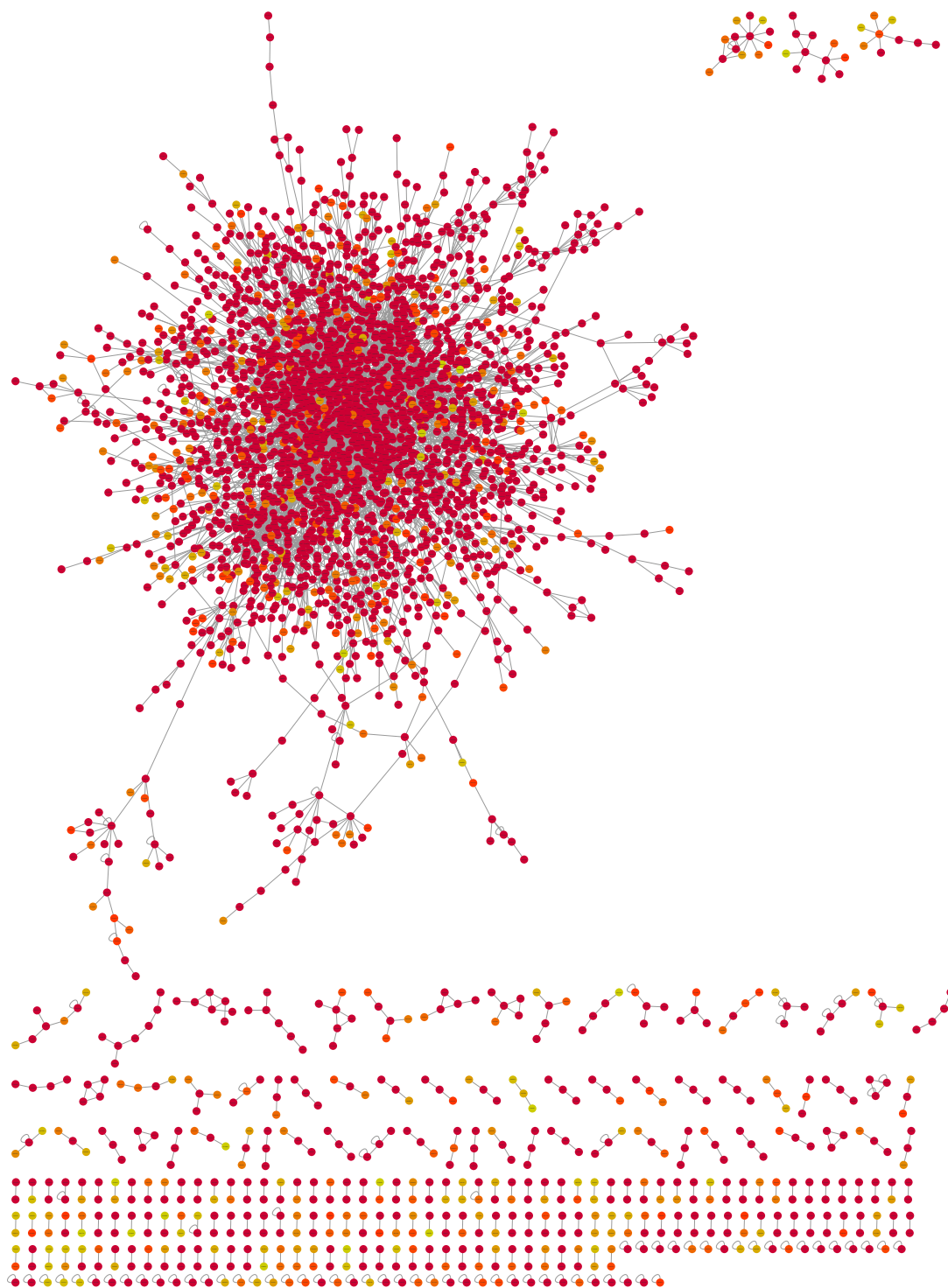


Figure 4.1: Network representation of a high confidence subset of the human interactome. As in [Figure 2.2](#), nodes indicate proteins, edges depict binary PPIs. Only interactions with a mentha⁵⁸ confidence score above 0.7 have been considered, reducing the network size from originally 14 650 proteins and 143 947 interactions to a subset of 5 175 interactions between 3 017 proteins. The network has been illustrated in Cytoscape 3.0.2 using a force-directed layout, maintaining the original color gradient from [Figure 2.2](#), where proteins with few interaction partners are colored yellow and those with more than ten partners are dark red. Compared to the unfiltered one, this network is split into more disconnected components.

alogs, that is, products of genes that resulted from inner-species gene duplication events⁸⁶.

4.1.2.4 *Functional annotation*

An observation that has early been made in interactome analysis is that biologically relevant interactions often involve proteins that have similar functional annotations^{20,215,367,382}. This fact has been employed by scoring methods that compute the functional similarity of protein interactors, for instance, using semantic similarity of Gene Ontology (GO) annotations^{176,187,291,343}. Functional similarity has been shown to be a good filter for high-confidence interactions^{93,215,249}. Biologically relevant protein interactions have been found between proteins with different annotation profiles³⁸⁸. In addition, GO, which is commonly used as a source for functional annotations, has been linked with significant annotation errors and biases due to its incompleteness^{57,385,397}.

4.1.2.5 *mRNA (co-)expression*

It has been shown that proteins that are involved in biologically relevant PPIs exhibit higher correlations in their expression profiles than non-interacting protein pairs¹⁴². This finding has been incorporated into different scoring approaches^{20,86,338,355,367} or was used to create tissue-specific interaction subnetworks²²⁶. Co-expression is a useful indicator for true biological interactions. [Bader et al.](#) have shown that co-complex proteins can show low co-expression values due to differences in protein degradation. For co-immunoprecipitation experiments, they even found an inverse correlation²⁰. In addition, the incompleteness of expression data can lead to biases as it generally favors proteins that have annotations over those without.

4.1.2.6 *Topological network measures*

The network representation of interactions, in which nodes correspond to proteins and connecting edges indicate different kinds of relationships between them (see [Figure 4.1](#)), allows for applying graph-theoretical analysis approaches^{22,23,183,392}. For example, using different measures of network topology^{16,304}, it has been shown that disease-associated networks show a higher connectivity than expected³⁰⁶ and that their modularity is altered³⁷³. Graph-theoretical network analysis measures are commonly used for comparisons^{125,226,313} and are the basis for various confidence scoring schemes, for instance, building on the local network neighborhood of a protein¹³⁹, by using clustering algorithms^{187,203}, or network transformations⁴¹⁶.

4.1.2.7 *Protein domains*

The interaction between proteins is often mediated by an underlying interaction of protein domains or short linear motifs^{286,287,320}. Based on this assumption, the

domain compositions of the protein interactors, for instance, their assigned Pfam domains¹¹⁸, can be probed for known domain-domain interactions (DDIs) in order to identify a subset of higher confidence^{3,198,217,244}. Such domain interactions can be determined based on different data types. The method considered to have the highest quality is the analysis of experimentally solved crystal structures that are stored in the Protein Data Bank (PDB³⁴). The databases 3did²⁵¹, iPfam¹¹⁶, and PInS⁴³ use this approach. In addition, a number of prediction algorithms have been proposed, which are based on the combination of heterogeneous data. Exemplary data types include protein interactions (potentially creating a circular argument if PPIs are used to predict DDIs, which in turn are used to validate PPIs), sequence information, or phylogenetic profiles^{227,261,415}. Integrating a large number of DDI datasets, Kim et al. were able to explain around 50 percent of the protein interactions by underlying DDIs¹⁹⁸.

4.1.2.8 Protein structure

Three-dimensional protein structures may be used to explain interactions between proteins on a molecular level, by identifying the particular protein binding interfaces⁹. For example, Aloy et al. have shown that interologous protein pairs tend to maintain their binding interfaces⁶, which allows for modeling homologous interaction pairs based on known structures²⁵⁰. Conversely, structural models may also be used to identify potential false positives in experimental data, for example, if proteins are reported to interact directly, although their structures make this interaction highly unlikely^{8,106}. The experimental determination of high resolution protein structures using X-ray crystallography, nuclear magnetic resonance spectroscopy, or electron microscopy is currently not amendable to high-throughput, with the result that the number of structures stored in the PDB grows slower than most other biological databases, requiring approaches for computational structure modeling^{250,402}.

4.1.2.9 Multiple data types

The incorporation of multiple data types is based on the underlying assumption that heterogeneous data can complement each other and account for potential problems and shortcomings of the individual datasets. The data types used by the different approaches are very diverse, including, but not limited to, all of the aforementioned proteomic, genomic, and functional datasets³⁶⁷. The combination is performed using Bayesian models or networks^{20,179,217,310,322,395,412} or machine learning approaches like probabilistic decision trees⁴²⁶. For all these integrative prediction or scoring techniques, the aforementioned gold standard sets are of critical importance^{52,178,420}.

4.2 Decentralized scoring systems

In the previous part of this chapter, I discussed the need for independent means for determining the quality of molecular interaction data. The methods that have been developed to this end are based on fundamentally different biological aspects of interactions, including their experimental detection systems, functional annotation of protein interactors, or evolutionary conservation. Given this diversity and the finding that no single scoring method is clearly superior³⁶⁷, a comprehensive scoring solution that incorporates all these approaches is not to be expected. In fact, it remains questionable whether it will ever exist²⁸⁰.

Centralized frameworks for protein confidence scoring Nevertheless, a number of computational platforms have been developed that use some of the aforementioned techniques for scoring interaction data, including EnrichNet¹³⁷, HIPPIE³³⁷, HitPredict²⁸⁴, INStruct²⁴⁴, IntAct¹⁹⁰, Interactome3D²⁵⁰, IntSore¹⁸⁷, PINA⁷⁷, PRINCESS²¹⁷, MINT²²⁰, and STRING¹²¹. From this selection, only Interactome3D, IntScore, PINA, and PRINCESS allow for scoring user-defined interaction data. The others are limited to pre-defined datasets, essentially making them (aggregate) interaction databases with added confidence scores (see [Section 2.3](#)). Interactome3D is largely based on protein structure prediction, IntSore, PRINCESS, and PINA all incorporate functional annotations and network topology, PRINCESS additionally makes use of evolutionary conservation and domain interactions^{77,187,217,250}.

These scoring platforms use a similar strategy as the aforementioned aggregate interaction databases, in the sense that they try to make a number of datasets or methods available at a single location. Since this is done by building central repositories, a major limitation inherent to these frameworks is that they cannot easily be changed or extended. The inclusion of a new scoring algorithm or an update of an existing method can only be accomplished by the central authority. Like in the previous [Chapter 3](#), this was one of the main arguments why we proposed solutions based on data decentralization, which will be presented in the remainder of this chapter.

4.2.1 Confidence scoring with DASMI

Our Distributed Annotation System for Molecular Interactions (DASMI) was designed for the exchange and the annotation of interaction data. [Section 3.2](#) primarily focused on the aspect of data availability, illustrating how scattered sets of interaction data are queried from and visualized at a single location. Here, I will describe how the same system can be used for the integration and visualization of different confidence scoring approaches.

[Figure 3.2](#) illustrates the DASMI architecture that consists of two server types: interaction servers providing the information that certain molecules interact and

confidence servers that make additional interaction annotation available in the form of confidence scores. Both server types use the same data exchange specification for requests and responses (see Section 3.2.1). The server installation tutorial on the DASMI website^a, for example, shows how the same interaction data may be visualized as a normal interaction source or as a quality measure.

4.2.1.1 DASMIweb visualization

A DASMI source that is registered with the label *interaction quality measure* will not appear in the DASMI interaction view (see Figure 3.6). Instead, it will be used to display additional information on top of the existing interaction data (see Figure 4.2). DASMIweb, for instance, will color the squares that represent particular interactions accordingly. Scores that have a limited range will be visualized with a color gradient, scores that cannot be transformed into a gradient will be highlighted, and all interactions for which there is no score will be greyed out. This visualization is also available for the confidence scores that are already provided with the respective original datasets, for example, the aforementioned IntAct or MINT-scores.

DASMIweb enables the confidence scoring of user-defined interaction data by allowing the upload of interactions in the form of PSI-MI XML 2.5 files. For further analysis in external applications like Cytoscape, the resulting scores are available for download along with the interaction data⁴¹.

4.2.1.2 DASMI confidence scoring servers

Our current setup consists of two DASMI sources that can be used for quality assessment of molecular interactions (blue rows in Figure 3.8):

- **FunSimMat** is based on the finding that interaction partners commonly share similar functions (see Section 4.1.2.4). For each protein in the UniProt database, FunSimMat precompiles and stores a number of similarity scores using different measures of semantic similarity^{340,341}. The FunSimMat DASMI source only uses simRel max scores developed by Schlicker and Albrecht, which are computed for each of the Gene Ontology (GO) categories: *BPscore* based on the biological process (BP) annotation, *CCscore* for cellular component (CC) terms, *MFscore* using molecular function (MF) annotations³⁴⁰. All these scores range between zero (no similarity) and one (identical annotation) and are presented by a color gradient if selected in DASMIweb.
- **Domain support** provides two measures that may help to explain protein interactions by underlying domain-domain interactions (DDIs) (see Section 4.1.2.7). To this end, it first retrieves the Pfam¹¹⁸ domain annotations for each protein interactor and tests for the existence of known DDIs. The score named

^a<http://dasmi.de/tutorial.pdf>

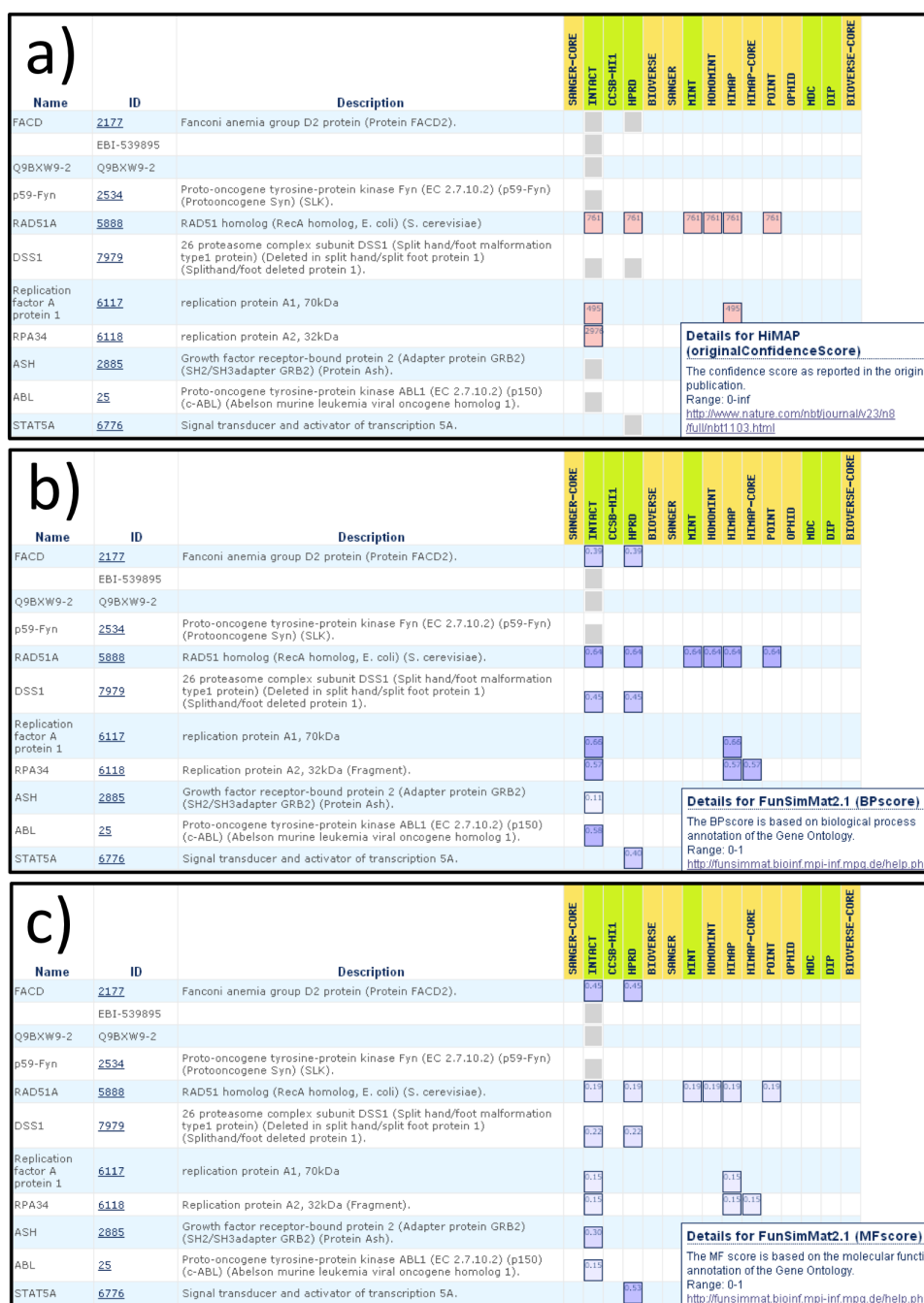


Figure 4.2: Visualization of confidence scores in DASMIweb. The selected confidence score, briefly described in the bottom right of a screen, is used to highlight interactions. Interactions for which no score is available are greyed out (see Figure 3.6). **a)** Confidence score provided by the Human Interactome Map (HIMAP). As this score has no limited range, all instances are marked in the same color. **b)** FunSimMat BPscore based on the biological process (BP) annotation of the Gene Ontology (GO). The zero-to-one range of the score is indicated by a color gradient, where white represents no similarity and dark blue indicates perfect similarity. **c)** FunSimMat MFscore based on GO molecular functions annotation (MF).

crystal-structure combines the DDI datasets 3did²⁵¹, iPfam¹¹⁶, and PInS⁴³, which have been extracted by analyzing experimentally determined crystal structures. The score named *predicted* combines several resources that use computational algorithms for predicting DDIs (see Table 3.3). Both scores are unary, indicating with a one those PPIs that might be explained by an underlying DDI. The DASMIweb interaction detail view (see Figure 3.7) lists the particular protein domains, their interactions to other domains, and potential scores provided by the DDI resource.

4.2.2 PSI Common Confidence Scoring System

DASMI showcased how distributed scoring methods may be made available to researchers at a single location. Therefore, after realizing that a single, comprehensive scoring solution is not within sight, the HUPO-PSI decided to develop a common scoring system following DASMI's decentralization strategy²⁸⁰. A particular requirement was that the scoring system should be based on the data exchange formats defined by the HUPO-PSI, namely PSI-MI XML and MITAB, instead of the DASINT format used by DASMI. Being responsible for DASMI, we emerged as the lead developers for this new framework named PSI Common Confidence Scoring System (PSISCORE)^{269,272}.

Figure 4.3 illustrates the general architecture and the basic workflow of PSISCORE. At the beginning, a HUPO-PSI document is loaded to a PSISCORE client. The document is then sent to user-selected scoring servers, which compute confidence values and add them to the input file. Each scoring server returns the input file with the added confidence scores back to the client. In a final step, all individual results are combined by the client into a single file, which is then returned to the user.

4.2.2.1 Web service definition

PSISCORE has been implemented as a web service accessible via the Simple Object Access Protocol (SOAP) (see Section 3.3). The definition of its methods, including arguments and return types, is done via a Web Service Description Language (WSDL) document. Being a community standard, the PSISCORE WSDL has been developed and discussed over the course of several HUPO-PSI meetings and workshops. Apart from our contribution as leading developers, Jules Kerssemakers provided major feedback and further comments were incorporated from Jens Hanssen, Javier De Las Rivas, and other workshop attendees.

The PSISCORE WSDL defines the following methods that a PSISCORE server has to provide, the full document is available in the Appendix:

- `getSupportedScoringMethods()` returns all scoring algorithms, which a server offers, as a list of `AlgorithmDescriptors`. An `algorithmDescriptor` specifies

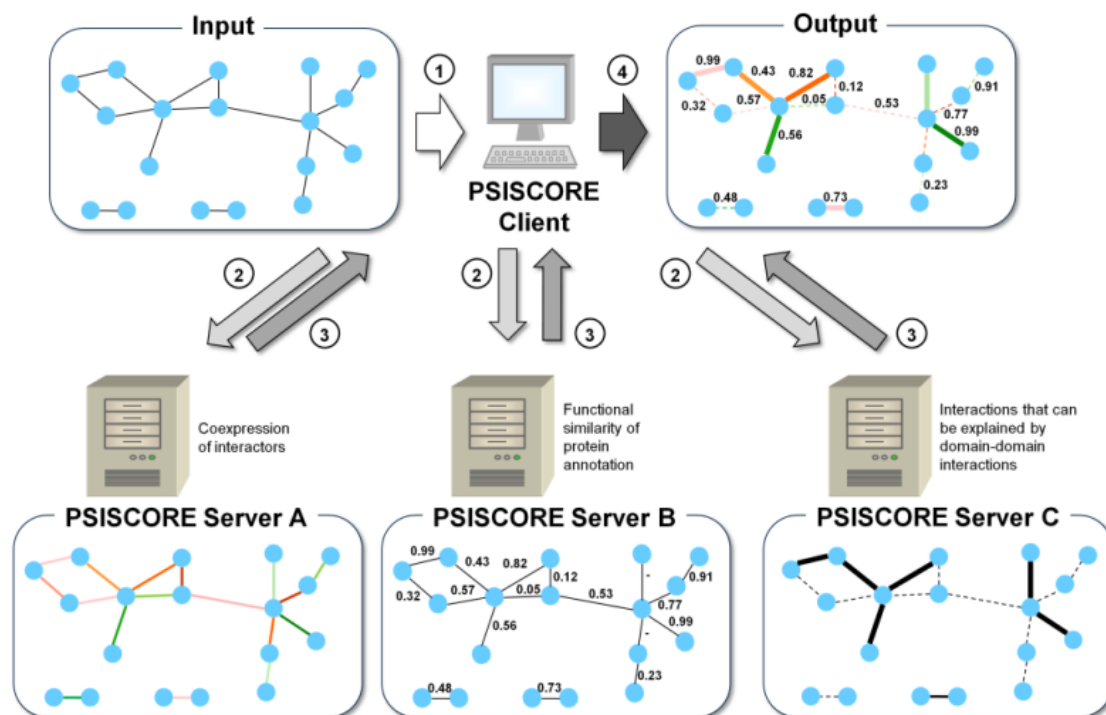


Figure 4.3: PSISCORE architecture and basic workflow. (1) A PSI-MI document is loaded to a PSISCORE client, which (2) sends it to several scoring servers. The servers perform their computations, add the confidence scores to the input document, and (3) return this document. (4) The client unifies the individual responses and provides the results to the user.

the name and type of a scoring algorithm (e.g. network- or structure-based), the range of its score (e.g. zero-to-one, zero-to-infinity), optional and required parameters, and data fields that need to be provided in the input file.

- `getSupportedDataTypes()` returns the input data types that a server can handle. At the moment, the two supported formats are PSI-MI XML and MITAB. For example, a server may require PSI-MI XML documents for scoring protein complexes or because the XML files contain some data fields that are unavailable in MITAB documents.
- `submitJob()` requests the server to score the interactions specified in the data section of the request. The `algorithmDescriptor` objects that have been retrieved in `getSupportedScoringMethods()` are used to select specific scoring algorithms and to provide the required parameters. The method returns a `JobResponse` object that contains the identifier of the scoring job and a suggestion after which duration to check if the job is completed. The rationale for this polling interval is that scoring algorithms that do a database query for pre-calculated scores have different computing times than scoring algorithms that do three-dimensional structure alignments or other computationally expensive

methods.

- `getJobStatus()` requests the status of the job with the corresponding identifier. The response is a status code that specifies whether it is still in the queue, currently being processed, or has finished with or without errors.
- `getJob()` retrieves the data associated with a certain job identifier after the scoring has finished. The response contains the scored interactions and a report that may provide a brief summary, warnings, or other information.
- `getVersion()` returns the version of the PSISCORE reference implementation that is used by the scoring server.

The PSISCORE WSDL also defines how the server behaves in unexpected or erroneous cases using exception objects like the specific `JobStillRunningException` and `InvalidArgumentException` or generic exceptions like `PsiscoreException`.

Computed scores are added to the respective sections of the input file, preserving potentially existing scores. For MITAB documents the scores are added to the confidence column, for PSI-MI XML documents a new `confidence` element is added to the `confidenceList` within the `interaction` element. In general, we encourage score providers to normalize their scores into a range from zero to one whenever possible.

4.2.2.2 Reference implementation

We developed open-source Java-based reference implementations for PSISCORE server and client that implement the methods specified in the aforementioned WSDL. The software is available as source code or precompiled packages along with further documentation from the PSISCORE project website^b. The core functionality is modeled after the PSICQUIC reference implementation^c that was provided by Bruno Aranda while he was working at the European Bioinformatics Institute.

Implementation details The central class of the server reference implementation is `DefaultPsiscoreService`. This class provides all the methods defined in the WSDL, maintains the scoring jobs and their mapping to particular score calculator threads, and unifies the result of each calculator into a final result that is returned to the client. Compared to PSICQUIC, where all the interactions are stored in a single MITAB file, the PSISCORE reference implementation has to be more versatile. While confidence scores might be provided from a MITAB file, they are also retrieved from database systems, or are computed on the fly, incorporating different elements from the input data.

^b<http://psiscore.googlecode.com/>

^c<http://psicquic.googlecode.com>

The common basis for all score calculators is defined by the `AbstractScoreCalculator`, which inherits basic threading capabilities from `Java Thread`. The `SimpleScoreCalculator` extends this class and provides most of the fundamental functionalities that are required for scoring, like converting the input data, regardless of whether they are described in MITAB or PSI-MI XML, to a common `EntrySet` representation and providing means for iterating through the interactions in this `EntrySets`.

The actual score calculator classes only need to extend the `SimpleScoreCalculator`, by implementing the two methods `getInteractionScores()`, which performs the scoring of individual interactions, and `getSupportedScoringMethods()`, which tells the `DefaultPsiscoreService` what scoring algorithms a calculator provides. The input to the `getInteractionScores()` method is an `Interaction` object of the `EntrySet` input representation, which contains references to all its participating interactors and other interaction details. The method will eventually add computed scores as new `Confidence` objects to the `Interaction` object.

On the client side, the basic methods and query capabilities are defined by the `AbstractPsiscoreClient`. Extending this class, the `SimplePsiscoreClient` provides all the functionalities to communicate with a particular PSISCORE server, including methods to submit jobs, get their status, or retrieve finished jobs. The `PsiscoreMetaClient` communicates with the registry and creates `SimplePsiscoreClient` objects for each of the scoring servers, thus allowing to query multiple scoring servers in parallel. It further provides methods to submit jobs to multiple scoring servers, keep track of the different job statuses and identifiers, and unifies the individual server responses once the scoring is completed.

The PSISCORE reference implementations of server and client use Apache Maven^d for managing project settings and dependencies. After downloading the reference implementation from the project website, a simple command is sufficient to download all required libraries, compile the source code, and package it into a Web Archive (WAR) file. This file can be placed in a Java servlet container like Apache Tomcat in order to make it publicly available over the internet.

4.2.2.3 Available scoring servers

Using our reference implementation, different PSISCORE servers have been set up at research institutions in Germany, Italy, Norway, and the United Kingdom. The following servers are listed in the PSISCORE registry in December 2013.

4.2.2.3.1 MPII The PSISCORE server at the Max Planck Institute for Informatics (MPII) provides five scoring algorithms. The three scores based on functional similarity (*BPscore*, *CCscore*, and *MFscore*) and the two scores based on domain

^d<http://maven.apache.org/>

interactions (*Domain support structure*, *Domain support inferred*) have already been described in [Section 4.2.1.2](#).

4.2.2.3.2 MINT The Molecular Interaction Database (MINT) offers its *MINT-score* and *HomoMINT-score* via a PSISCORE server. Both scores take into account experimental evidence associated with the interaction (see [Section 4.1.2.1](#)) and range between zero and one, where one indicates the highest confidence in an interaction. The scores are pre-calculated and are thus only available for interactions that are reported in the MINT databases. Detailed information on the scores can be found in the respective publications^{65,290} and online^e.

4.2.2.3.3 MI score The molecular interaction (MI) score is similar to the MINT-score in the sense that it is also calculated based on the number of publications, the experimental detection methods, and the interaction types that support a certain interaction. The score ranges between zero and one, where one indicates the highest confidence in an interactions. Four different MI scores may be computed by the PSISCORE server on-the-fly, based on interaction evidence from different sources: *MI score - intact* only uses data from the IntAct PSICQUIC server, *MI score - psicquic sources* uses all PSISCQUIC servers that are available at the PSICQUIC registry, *MI score - intact plus imex curation* combines results from IntAct with results from other IMEx servers, and *MI score - imex curation* uses only IMEx databases. Further information about the score is available at the project website^f.

4.2.2.3.4 iRefIndex iRefIndex provides three scores computed from the publications that provide evidence for an interaction³¹⁵. *lpr* (lowest PubMed identifier re-use) is the lowest number of interactions that a publication, which provides evidence for a particular interaction, is used to support²⁴⁷. Low values indicate that at least one of the publications that support an interaction has reported only few interactions, which is likely the case for low-throughput experiments. *hpr* (highest PubMed identifier re-use) is the highest number of interactions that a publication supporting an interaction is used to support²⁴⁷. High values indicate that a publication describes many other interactions, which is typical for high-throughput studies. *np* (number PubMed identifiers) is the total number of unique publications that support the interaction. Higher values indicate that an interaction has been reported in multiple publications. All three scores are not normalized and range from zero to infinity.

^e<http://mint.bio.uniroma2.it/mint/doc/MINT-confidence-score.html>

^f<http://miscore.googlecode.com/>

4.2.2.4 *PSISCORE registry*

The PSISCORE registry^g maintains a list of all available scoring servers and the algorithms they provide. An early prototype of the service has been developed by Sascha Meiers during an internship in our group at the MPII. The registry is implemented as a REST-based service (see Section 3.3) that is queried via specific URLs. The following query parameters are supported, their names and values have been chosen to correspond to the PSICQUIC registry wherever possible:

- **action** may be used to select all servers or only those that are active or inactive using the corresponding values **STATUS**, **ACTIVE**, and **INACTIVE**, respectively.
- **format** defines how the registry presents the results. If it is not provided, a human-readable tabular HTML document is generated. To obtain the same detailed data in a format that is more suitable for computational processing, a request with the value **XML** can be made. The value **TXT** will output a plain textual representation consisting only of the server name and its URL.
- **name** may be used to limit the query to a particular scoring server, for instance, to retrieve details on its scoring methods. If no scoring server with the name exists, the registry will not show any results.

The following examples show different combinations of these parameters. The URL <http://psiscore.bioinf.mpi-inf.mpg.de/registry?action=active&format=txt> returns all scoring servers that are currently active in a plain text format, consisting of a unique ID and the SOAP URL. The query URL <http://psiscore.bioinf.mpi-inf.mpg.de/registry?name=miscore&format=xml> returns a detailed description of the scoring server with the name "MI score" as an XML document. In addition to the server name and URL, the document also contains a brief description of every scoring method the server provides, including a textual description, keywords, and the range of the score. The same information would be represented in a human-readable form by removing `&format=xml` from the query.

4.2.2.5 *PSISCOREweb*

Based on the client reference implementation, we developed PSISCOREweb as a simple PSISCORE client to show the potential of decentralized scoring. PSISCOREweb has not been developed for large-scale analyses like the annotation of whole interactomes. These tasks require downloading the client library from the PSISCORE project website and running a client locally.

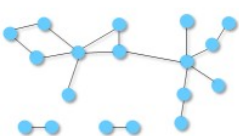
Scoring interactions with PSISCOREweb essentially consists of four steps:

1. uploading the interaction data to PSISCOREweb,


^g<http://psiscore.bioinf.mpi-inf.mpg.de/registry>

Score your interaction data

1) Upload data.
Upload your interaction data in the form of a valid PSI-MI file.

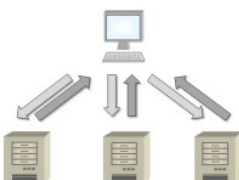


2) Select scoring methods.
Select the scoring methods you want to run on your data. You can get more information on each method by hovering the mouse over its name.



- <http://mint.bio.uniroma2.it/psiscore-ws-0.9.7-SNAPSHOT/webservices/psiscore>
 - MINT-score
 - HomoMINT-score
- <http://biotin.uio.no:8081/psiscore-ws/webservices/psiscore>
 - np-number_pmids
 - lpr-lowest_pmids_reuse
 - hpr-highest_pmids_reuse
- <http://psiscore.bioinf.mpi-inf.mpg.de/psiscorews/webservices/psiscore>
 - BPScore
 - CCScore
 - MFScore
 - Domain support, inferred
 - Domain support, structural
- <http://www.ebi.ac.uk/enfin-srv/miscore/webservices/psiscore>
 - MIscore - intact
 - MIscore - psicquic sources
 - MIscore - intact plus imex curation
 - MIscore - imex curation

3) Start scoring.
Press the button to the right to send your data to the selected scoring servers. The confidence scores will be included into your input file as soon as they are calculated.



4) Download results.
Download your PSI-MI file with all confidence scores by pressing the button to the right.

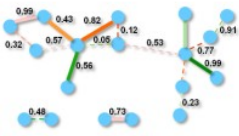


Figure 4.4: The user interface of PSISCOREweb is split into four parts that are activated subsequently. First, the interaction data are uploaded to the client as a valid PSI-MI document. Then, the scoring servers and their methods are selected and the scoring process is started. In a last step, the user may check if the scoring jobs have finished and retrieve the resulting file via a personal download link.

2. selecting the desired scoring methods,
3. sending the interaction data to all selected scoring servers,
4. retrieving the resulting scored interactions.

Figure 4.4 shows the web interface, where the user is guided through these steps. The start and end point in PSISCOREweb are valid PSI-MI documents, either PSI-MI XML2.5 or MITAB2.5 files. The initial file is selected and uploaded to the server. If it can be parsed successfully, a status message is written and the next block in the user interface is activated and marked by a strong blue border. The user may then select the scoring methods that should be incorporated from each server. The selection here is depending on the servers that are currently available and active in the PSISCORE registry. By default, all scoring methods are selected. By clicking on the respective button in the section below, the scoring process is then started. Each server will perform the requested scoring computations and include the resulting scores into the input file. By clicking on the request job button in the final section, PSISCOREweb will retrieve the individual results if all jobs have finished. The scores are then extracted from the individual response files and inserted into the user input file, leaving the rest of the file content unchanged. The resulting file can be downloaded using a personalized link.

4.3 Conclusions

4.3.1 Summary

Molecular interactions are of fundamental importance for understanding the function of complex systems like the cell. Consequently, more and more research groups are working in this area, producing an ever-increasing amount of data on the interaction of molecules. However, the experimental detection techniques or computational prediction algorithms, with which these interactions may be determined, differ greatly, as each of them has certain strengths and weaknesses. Skepticism and quality issues have been associated with particular approaches for more than a decade, leading to a state where independent means to assess individual interactions are required, regardless of whether they have been determined by high-throughput screens, small-scale single-protein studies, or computational predictions.

In response to this, a number of frameworks have been developed for scoring interaction data quality^{77,187,217,250}. The measures that have been proposed to this end are based on different biological aspects of an interaction, including structural information, network topology, experimental conditions, similarity of the functional annotations of the interacting molecules, or evolutionary conservation. It is thus practically unfeasible that a single, central scoring approach could combine all these

methods²⁸⁰. Therefore, we proposed a decentralized approach, where multiple scoring servers are hosted at different locations. An instant benefit of this decentralization is that such systems can easily be extended if new aspects of molecular interactions are to be scored. In addition, decentralization enables research groups to focus on their particular strengths. For example, a group with a strong background on structure modeling may develop a structure-based algorithm, groups with expertise in functional similarity focus only on this aspect, while usability experts and designers may develop useful and aesthetically pleasing graphical user interfaces.

We first showcased the potential of decentralization by using the Distributed Annotation System for Molecular Interactions (DASMI) for confidence scoring. Although the main focus of DASMI is on the exchange and annotation of interaction data, it proved to be useful for integrating and visualizing information about interaction data quality, as all DASMI servers use the same data exchange specification. We set up two scoring servers that may assess the functional similarity of the interacting molecules based on all three categories of the Gene Ontology and that may explain protein interactions by underlying domain interactions. The DASMI client DASMIweb can incorporate these and other scoring algorithms and visualize them with color gradients that promote visual inspection of interaction data. In addition, DASMIweb enables the scoring of user-defined interaction data and allows to download the results for further external analyses⁴¹.

After realizing that decentralization is a key strategy for interaction confidence scoring, the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) decided to implement a common confidence scoring system modeled after DASMI²⁸⁰. We developed this system named PSI Common Confidence Scoring System (PSISCORE) based on the HUPO-PSI document formats PSI-MI XML and MITAB. In collaboration with other PSI members, we designed the web service definition, developed reference implementations for servers and clients, and used these to set up different scorings services¹⁴. In addition to the aforementioned servers providing functional similarity and domain interaction support scores, PSISCORE servers have been set up in other countries to score interactions based on the experimental evidence or the publications that support them^{220,247}. In addition, we developed the exemplary client PSISCOREweb, which may be used to upload an interaction data file, send this file to multiple scoring servers in parallel, retrieve and unify the individual scoring results, and provide the final result back to the user.¹⁴

4.3.2 Outlook

Distributed systems have shown their usefulness for interaction data exchange, annotation, and confidence scoring. However, the three examples mentioned throughout this chapter, the Distributed Annotation System for Molecular Interactions (DASMI), the Proteomics Standards Initiative Common Query Interface (PSIC-QUIC), and the Proteomics Standards Initiative Common Confidence Scoring System (PSISCORE), have also shown how such systems critically depend on constant

support if they are to be largely supported. PSICQUIC, which is maintained by several full-time software developers, has achieved a critical size⁸⁸. DASMI and PSISCORE have acquired a number of users, but are both lacking greater public support on the server side. In order to increase the number of external providers of PSISCORE servers, additional efforts are required to convince and constantly support interested service providers.

With PSISCORE being the official successor to DASMI, all future efforts should be focused on the former. The reference implementations would benefit from optimizations to increase the overall performance and further improve the usability. In addition, PSISCOREweb might be turned into a more versatile scoring platform by removing the current file size limitation and incorporating visualizations of the interaction data and the corresponding confidence scores. Developing additional PSISCORE clients, for example, for Cytoscape, Galaxy, or the R programming language, would easily extend the user base of PSISCORE and make the system available for further analysis workflows.

Thus far, the primary focus of PSISCORE has been on the exchange of interaction confidence measures. Computed confidence scores were simply added to the input data. While it remains questionable whether comprehensive scoring solutions will ever be practical, considering the large variety of different scoring approaches and the contradicting aspects of molecular interaction that can be assessed, the ultimate goal is to combine the different scoring schemes in a useful manner. To this end, further research is needed to develop approaches that combine various scoring algorithms in a sophisticated manner.

5 Integrative network analyses of viral host factor screens

This chapter describes applications of different biomedical data integration and analysis techniques to investigate specific biological questions. In particular, we focus on the analysis of experimental screens that aim to determine human host factors required for viral infections. In [Section 5.1](#), an introduction into viruses like the hepatitis C virus (HCV) is provided, followed by a description of the experimental systems to determine human host factors. In [Section 5.2](#), the computational framework that we have implemented for the bioinformatics analysis of the resulting screen data is presented. This section also contains a functional network analysis of published HCV host factor screens. Our framework is then used to analyze and prioritize two independent large-scale screens determining host factors for HCV; the results are presented in [Section 5.3](#). I will close the chapter by summarizing our contributions and providing an outlook on future work.

5.1 Introduction

Viruses and the infections they cause pose a major threat to human health. Despite intense research efforts, the therapy of medically important infections, such as caused by the human immunodeficiency virus (HIV), Dengue, influenza, or the hepatitis C virus (HCV), remains a challenging and partly unsolved problem²⁹⁷. The situation is particularly problematic in developing countries, where knowledge about certain infections is not widespread, the standards of medical care are lower than in the industrial nations, and the climatic conditions may promote some viral transmitters like mosquitos.

5.1.1 Hepatitis C virus

About 180 million people worldwide are chronically infected by the hepatitis C virus (HCV), which is central to our studies and the rest of this chapter³⁷⁶. HCV is a blood-borne virus that infects human hepatocytes³²⁶. Between 10% and 25% of the people that get infected manage to clear the virus in the acute infection stage within the first six month, the remaining 75% to 90% develop a chronic infection in their

liver¹⁴¹. The reasons why some patients manage to clear the virus, while others acquire a chronic infections, are not yet fully understood^{158,376}.

HCV is also called a "silent killer", as the infection remains largely unnoticed in the first years¹⁰⁰. Over time, however, the permanent infection and inflammation of the liver may lead to steatosis, also known as fatty liver, and functionless scar tissue, resulting in liver fibrosis or cirrhosis¹³⁵. Ultimately, chronic hepatitis may lead to hepatocellular carcinoma, a form of liver cancer³⁷². In Europe and the United States of America, HCV is the leading cause for liver transplantation³⁷².

Treatment options Permanent eradication of the virus, measured as sustained virologic response (SVR) 24 weeks after completion of the treatment, can be achieved in a large number of patients²⁰⁷. Until recently, the recommended treatment regimen for chronically infected patients was a combination of the immune-system stimulating interferon and the antiviral agent ribavirin^{157,361}. Depending on the viral genotype, this treatment has a success rate of 5-80%²⁹⁷. However, it frequently leads to undesired and potentially strong side-effects that may result in patients aborting the 48-week long treatment¹³¹. In 2011, new direct-acting antiviral agents (DAAs), which target non-structural HCV proteins like the NS3 protease, have been approved for therapy⁹¹. This opened up new treatment options, as DAAs can be used in shorter treatments combinations, thus leading to less side-effects. In addition, they generally achieve better SVR rates than the classic treatments^{390,425}. However, DAAs may lead to the rapid development of resistance mutations in the viral genome⁴⁰⁴, which is promoted by the error-prone viral polymerase³³¹. The rapid turnover, caused by the short lifespan of HCV particles⁸¹, may thus lead to viral populations that are drug-resistant. With new DAAs in the pharmaceutical drug development pipelines^{30,87,283}, individual, patient-specific drug combination therapies, as they are widely used in HIV treatment²¹⁶, are required⁸⁵.

HCV genome and proteome HCV was discovered and cloned in 1989 as "non-A, non-B hepatitis"⁷¹. It is an enveloped, single-stranded RNA virus that is classified into the *hepacivirus* genus in the *flaviviridae* family⁴⁰⁵. Other flaviviruses, which share some commonalities in their life cycle, include Dengue or the yellow fever virus¹¹³. HCV's genome only has a length of about 9 600 nucleotides and encodes a single polyprotein that is post-translationally cleaved into three structural and seven nonstructural proteins²⁴⁸. The structural proteins core and the envelope glycoproteins E1 and E2 are required to form the viral particle. The nonstructural proteins p7, NS2-3, NS3, NS4A, NS4B, NS5A, and NS5B have diverse functions, for instance, as protease (NS3) or polymerase (NS5B)²⁹⁷. A detailed description of the molecular biology of HCV is, for example, provided in Brass et al.⁴⁷ or Tang and Grisé³⁷².

Host factors Viral genomes are very short and encode only a fraction of the genes that are required during their life cycle²⁹³. Consequently, viruses like HCV, which

only encodes 10 proteins, depend on the human host cell and numerous molecules like proteins, lipids, nucleic acids, or membrane compartments^{29,90,372}. These human molecules are known as *cellular factors* or *host factors*^{202,254}. An exemplary genetic host factor for HCV is a polymorphism upstream of the IL28B gene, which in genome-wide association studies has been linked with a significant difference in response to drug treatment and clearance of viral particles^{131,366,371,376}. Another well-known host factor is the liver-specific microRNA mir-122^{124,156}. In primates, constant degradation of this essential HCV host factor led to a long-lasting suppression of HCV infection, without measurable viral resistance mutations or noticeable side-effects²¹⁰.

In the following, we will use the term *host factor* only for referring to human genes and their proteins that influence the viral life cycle. Depending on their particular effect on the virus, host factors can be divided into two classes. *Host dependency factors* (HDFs) are proteins that are essential for the virus, for example, cellular receptors that mediate viral entry into human hepatocytes^{28,334}. *Host restriction factors* (HDRs) describe proteins that have an oppressing effect and control or restrict viral replication, for example, as part of the innate immune response⁹⁷.

Viral life cycle The life cycle of HCV can be broadly split into four stages (see Figure 5.1):

1. **entry** into hepatocytes,
2. **replication** in the cytoplasm,
3. **assembly** of viral particles,
4. **release** of infectious particles.

Two particular characteristics of the HCV life cycle are its strong connections to the host lipid metabolism^{10,245,368} and the formation of cellular structures know as *membranous web*, formations derived from host membranes that are home to viral replication complexes³⁷². A detailed description of the individual stages can be found elsewhere^{27,248,254,372}.

Generally, the two early stages, entry and replication, are better studied and understood than the latter, assembly and release (see Table 5.1). A reason for this is that for many years HCV was very difficult to study *in vitro*, as cultured viruses did not produce enough infectious viral particles to sustain a permanent infection²⁶. A major development in this direction was made by subgenomic replicons, self-amplifying RNA-based systems in hepatoma cells⁴⁰⁸. These artificial systems allow studying intracellular parts of the viral replication cycle like RNA amplification, but do not produce infectious viral particles. A breakthrough in the study of HCV was made by the discovery of a particular HCV genome in a Japanese patient¹⁸⁸, from which clones could be derived that have very high production rates of infectious

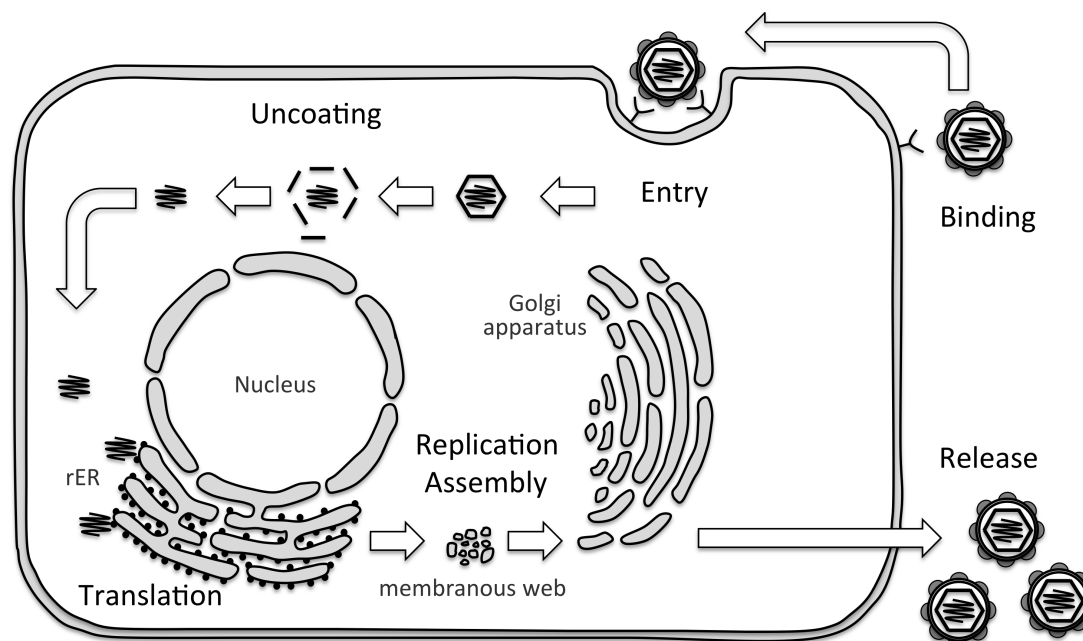


Figure 5.1: Life cycle of the hepatitis C virus (HCV). First, viral particles bind to specific receptor molecules on the cell surface. The virus-receptor complex then moves to tight junctions and enters the cell by means of receptor-mediated endocytosis. In an uncoating step, the viral nucleocapsid is then released into the cytoplasm and is relocated to the rough endoplasmic reticulum (rER), where translation of viral RNA occurs. The translated polyprotein is post-translationally modified and viral particles are produced within the membranous web. The viral life cycle is complete when particles exit the host cell through the secretory pathway. Image modified from [Poenisch and Bartenschlager²⁹⁶](#).

particles³⁹⁸. From 2005 onwards, such HCV cell culture (HCVcc) systems enable studies of the whole viral life cycle and the human host factors involved in it.

5.1.2 Host factor determination using RNA interference

Different experimental techniques exist for studying connections between viruses and their human hosts. RNA interference is the approach that is used in most of the screens that determine host factors for viral infections. In the next sections, the mechanism of RNA interference will be briefly introduced, followed by a description how this can be used in large-scale experiments.

RNA interference RNA interference (RNAi) is a mechanism of post-transcriptional regulation or gene silencing. Based on studies that exemplified how RNA fragments with antisense sequence to a target gene can effect its expression, [Fire et al.](#) provided the first model of this gene-regulatory mechanism in 1998¹¹⁹. For this discovery, Fire and Mello were awarded the 2006 Nobel Prize in Physiology or Medicine. In their experiments, they found that after injecting double-stranded RNA (dsRNA) into the worm *C. elegans*, genes containing a homologous sequence were silenced very

potently. The inhibition was much stronger than when single-stranded antisense RNA was introduced, suggesting an underlying cellular pathway for RNA processing. Elbashir *et al.* later showed that the effectors of RNAi are single-stranded short interfering RNA (siRNA) fragments with a length of 21 or 22 nucleotides¹⁰³. In the same year, Elbashir *et al.* proved that RNAi can be triggered in mammalian cell lines by introducing siRNA duplexes of 21 nucleotides length into the cell¹⁰². This discovery turned RNAi into a powerful technology for studying individual human gene function via the introduction of siRNAs that are specifically tailored to the mRNAs of certain target genes.

In the RNAi pathway, which is triggered after dsRNA is introduced into a cell, the two RNA strands are first processed by an RNase named Dicer³⁵. The guide strand, which is sequence complementary to the target mRNA, is incorporated into a multi-protein complex named RNA-induced silencing complex (RISC)¹⁴⁴. By binding to the target mRNA in a sequence-specific manner, RISC then leads to its degradation and thus to the temporary silencing of the gene expression. Recently, Lima *et al.* have shown that single-stranded siRNA (ss-siRNA) can be modified to exhibit similar silencing potency than the ds-RNA that is processed in the RNAi pathway²²³. Combining the simplicity of siRNA silencing with the efficiency of the RNAi machinery, Yu *et al.* were able to use ss-siRNA to silence mutant huntingtin in Huntington disease patients⁴¹⁷.

Loss-of-function studies using RNAi The power of RNAi for loss-of-function studies of human genes has resulted in the availability of commercial libraries that target all known or predicted human gene transcripts²⁴⁶. This has resulted in a great number of RNAi screens investigating fundamentally different processes like endocytosis²⁸⁸, signaling pathways^{122,246}, or host factors for bacterial¹ or viral infections^{70,154}.

The particular experimental setups vary between individual RNAi screens, but most of them follow the same general principles (see Figure 5.2). One approach of introducing siRNAs into cells is named reverse transfection¹⁰⁸. Specific siRNAs and a transfection reagent that promotes cellular uptake of siRNAs are spotted on a plate and a defined number of cells along with a nutrition medium are seeded on each of the spots. This starts the cellular siRNA uptake and results in the temporary silencing of the target genes⁴⁶. After a certain duration, the silencing effect can be determined using different readout methods. Depending on the particular biological question and the experimental setup, a reporter molecule may have been stained with a color marker beforehand, allowing the use of immunofluorescence microscopy for counting or measuring signal intensities of the reporter²³⁸. In a statistical post-processing step, these values are then normalized for variability within and between experimental repetitions or with respect to negative and positive controls that should show no or a strong effect, respectively²³¹. Commonly, the effect of a particular siRNA is quantified as a *z-score*, which is defined as the difference between the intensity value for a spot and the mean value of a plate, divided by the standard

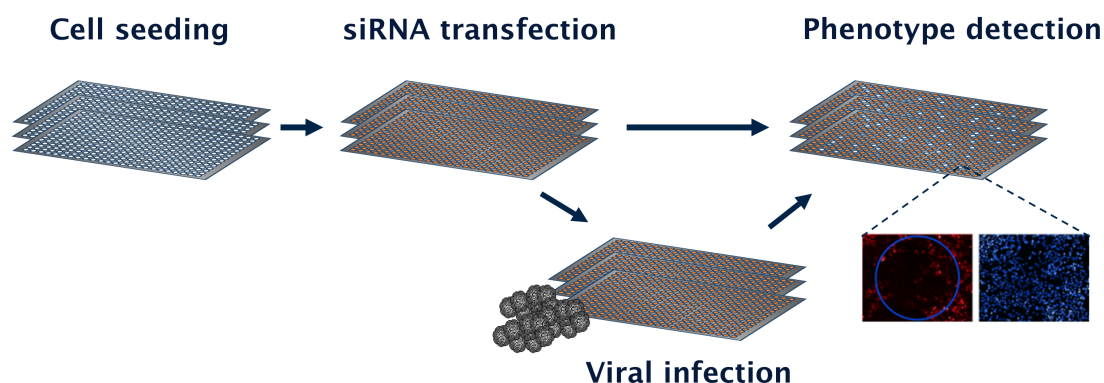


Figure 5.2: Schematic experimental setup of large-scale RNA interference screens¹⁰⁵. First, cells are seeded into the wells of a plate. In each well, cells are then transfected with siRNAs, leading to temporary silencing of the target gene expression. If the RNAi screen is used to detect host factors for viral infections, viral particles are added in the next step and given some time for infection and replication. After a certain duration, phenotypic changes can be measured. In case of viral host factor screens, this can be achieved by measuring viral particles and cells with multichannel fluorescence microscopy.

deviation of the plate³²³. Using statistical tests, a *p-value* may additionally be computed to assess the significance of the computed score⁴²⁸. Further information about different aspects of statistical post-processing including quality control, data normalization, and hit calling are provided in Amberkar et al.¹¹.

RNA interference screens for determining viral host factors are similar to the normal screen setup described above. The essential difference is that, after the cells have been spotted on the plate and the siRNA has been transfected into the cell, leading to degradation of the target mRNA, the cells are infected with viral particles³²¹. Depending on the particular viral study system, such particles may be complete infectious viruses or engineered pseudo-particles that only consist of viral envelope proteins. The former would allow studying the complete viral life cycle²¹⁸, while the latter may be used to detect host factors required for viral cell entry³⁹⁴.

Side-effects RNAi screens are faced with several challenges. First, the efficiency of different siRNAs varies strongly, depending on their chemical properties, their sequence composition, or the accessibility of the target binding site¹⁶⁰. In addition, side-effects or off-target effects of siRNA injections are very common.

Unspecific side-effects, which may occur independent of the particular siRNAs that are being used, may be triggered by the activation of the interferon system and thus the stimulation of an immune response³⁵¹. In addition, the transfection reagent induces stress on the cells, thus changing normal cell viability. Snijder et al. have further found that the studied cellular populations may be in different cellular developmental stages and thus show a variable response to the same treatment³⁵³.

Specific side-effects may occur when the siRNA does not only bind to the intended

mRNA, but has additional complementary targets in the transcriptome¹⁷⁵. Since siRNAs only have a length between 19 and 23 nucleotides, such complementary sequences appear frequently¹⁶⁰. In particular, siRNA binding is largely driven by a seed region comprising only 2-8 nucleotides, which further increases the number of potential binding sites^{36,174}. In addition to complementary binding, siRNAs have also been found to bind to targets that have imperfect sequence complementarity¹⁷³.

Several of the side-effects may be minimized by sophisticated design of the siRNA, optimizing its chemical properties in order to maximize the efficiency, while at the same time minimizing the complementary target sequences^{96,235,255,265,422}. In addition, it is crucial that for each gene that is tested, multiple different siRNAs are used such that each siRNA binds to a different sequence in the target mRNA³⁴⁷. Multiple replications of the same experiment may further account for effects of cellular variability and produce results that are statistically more robust and relevant. In summary, the screen design largely influences the quality of the resulting findings.

5.2 Computational framework for analyzing viral RNAi screens

In order to analyze and prioritize the results of large-scale RNA interference screens, in particular those determining host factors for viral infections, we built an integrative data warehouse combining multiple heterogeneous datasets.

5.2.1 Datasets

Published host factor screens Of central importance to our analysis are results from previously published host factor screens, as they allow placing individual results into a broader context^{70,154}. Table 5.1 lists all RNAi screens for host factors of the hepatitis C virus that we have collected into our database in a laborious manual curation process. This unpublished, in-house database of host factors turned out to be a very valuable resource, allowing us to perform broader analyses and functional studies. Most of the data it contains is not accessible in RNAi databases like GenomeRNAi³⁴⁴ or VIRsiRNAdb³⁷⁴.

Although several of the screens have been carried out using similar viral study systems, cell types, and RNA interference protocols, their results are largely complementary and show very little overlap. This has also been found in previous studies for HCV and other viruses^{70,430}. For example, between four independent influenza screens, the average pairwise screen overlap was less than seven percent¹⁴⁶. A similar value was found when comparing nine different HIV screens⁵⁵. This situation is thus similar to large-scale yeast-two hybrid screens for determining protein-protein interactions, which also had been questioned after showing very little overlap in their results (see Section 4.1).

Table 5.1: Published host factor screens for hepatitis C virus (HCV) stored in our database. The parts of the viral life cycle that were studied in the screen and the number of host dependency (# HDF) and host restriction factors (# HRF) that have been reported in the respective publication are listed.

Reference	Year	Parts of viral life cycle	# HDF	# HRF
Berger et al. ³³	2009	entry, replication	7	-
Borawski et al. ⁴²	2009	replication	108	38
Chao et al. ⁶³	2012	replication	1	-
Chen et al. ⁶⁹	2010	replication	1	-
Coller et al. ⁷⁴	2009	entry	16	-
Coller et al. ⁷⁵	2012	release	14	-
Hara et al. ¹⁴⁷	2009	replication	9	-
Herker et al. ¹⁵¹	2010	assembly	1	-
Jones et al. ¹⁸⁴	2010	replication	7	7
Li et al. ²¹⁸	2009	entry, replication, assembly, release	237	25
Lupberger et al. ²²⁸	2011	entry	58	-
Ng et al. ²⁶²	2007	replication	9	-
Randall et al. ³¹⁴	2007	replication	26	1
Reiss et al. ³²¹	2011	entry, replication	13	-
Supekova et al. ³⁶⁵	2008	replication	3	-
Tai et al. ³⁷⁰	2009	replication	96	-
Trotard et al. ³⁸⁰	2009	entry, replication	5	2
Vaillancourt et al. ³⁸³	2009	replication	73	-
Xue et al. ⁴¹³	2007	replication	3	-

Hao *et al.* have recently shown that, although screen overlaps in terms of genes or proteins are very modest, the screens detect host factors with similar functions¹⁴⁶. The overlap between functional categories that are found is 19%. In addition, the different result lists are significantly more connected in a protein interaction network than would be expected by chance¹⁴⁶. Based on their statistical model, they estimated that the low screen overlap is mostly due false negatives, host factors that are not detected in a screen, and not due to false positives as they might result from RNAi off-target effects¹⁴⁶.

Molecular interaction data In addition to information about previously published host factors, molecular interaction data are another fundamental resource for our analyses. To this end, we combine integrated networks of human protein-protein interactions and co-complex associations with information about interactions between human and viral proteins. The latter dataset largely consists of large-scale yeast-two-hybrid experiments by de Chasse *et al.*^{84,257}. We used different sources for human protein interaction data, consisting of aggregated data from BioMyn^{311,312}, STRING³⁶⁹, and iRefIndex³¹⁵.

Functional annotations BioMyn^{311,312} has also been used as resource for performing functional annotations and for computing functional enrichments, either incorporating Gene Ontology (GO) or pathway annotations³²¹. In more recent studies, we also used topGO⁴, which, compared to classic enrichment computations, also accounts for the graph structure of GO, thereby reducing redundancy in the results. Functional similarities between genes or proteins were retrieved from FunSimMat^{340,341}.

In addition, we used different online services in order to perform specific analysis tasks. For example, our recent studies used DAVID to annotate gene sets with UniProt keywords³⁷⁵ and functional GO annotations (see Figure 5.6). While this functionality in principle is also available via BioMyn, the latter resource was not available during the time of our study.

5.2.2 Functional network analysis of viral host factors

Several studies have investigated the molecular interaction networks involved in different viral infections, mostly focusing on topological network parameters. For example, Navratil *et al.* found that viral proteins particularly interact with highly central proteins and those having a high bridging centrality²⁵⁶. Hao *et al.* found that while individual screen results do not tend to form network clusters, the combined set of host factors tends to be highly connected¹⁴⁶.

Together with Nora Speicher, we developed a new network-based method to analyze sets of host factor candidates by combining them with molecular interaction data and functional protein and gene annotations³⁵⁴.

Materials and methods Based on the hypothesis that viruses require host factors with defined biological functions in order to fulfill their life cycle (see [Section 5.1.1](#)), host factor proteins may be assumed to be functionally more similar to each other than randomly chosen human protein pairs. In the following, two host factor candidates form a pair if their distance in a molecular interaction network is below a particular threshold. Here, the network distance of two nodes is measured by the shortest path length between these nodes³⁰³. Briefly summarized, our approach identifies all pairs of host factor candidates within a particular input set and compares the functional similarities of these pairs to the functional similarities of randomly generated protein pairs that lie within the same network distance.

We tested different measures for assessing functional similarity³⁵⁴ and found that sim_{Rel} , developed by [Schlicker et al.](#)³⁴², performed best for computing functional gene and protein similarities based on the available Gene Ontology (GO) annotations¹⁵. We used two integrative datasets as sources for molecular interactions, STRING³⁶⁹ and BioMyn³¹¹, in order to explore how the type of interaction data influences the resulting functional similarities. STRING contains weighted interactions where the weight reflects the strength of the combined interaction evidence that is derived from different types of direct physical interactions and functional associations. In order to exclude spurious interactions, we filtered the network by requiring a weight of at least 0.4. From the BioMyn database, we used all direct physical protein-protein interactions.

For our analyses, we selected the host factor determined by [Borawski et al.](#)⁴², [Li et al.](#)²¹⁸, and [Tai et al.](#)³⁷⁰ from all published HCV host factor screens (see [Table 5.1](#)). From these, we derived the two additional sets *AllScreens*, a union of the host factors that have been detected in at least one of the screens, and *HighConfidenceHits*, containing host host factors that have been found in at least two different screens.

Results After testing our method with artificial protein pairs taken from known protein complexes³⁵⁴, we applied it to the aforementioned sets of HCV host factors using different network distances and functional similarity measures. [Figure 5.3](#) exemplarily shows that pairs of host factors derived from the published HCV host factor sets and those from two additional datasets have higher functional similarities than randomly generated pairs. Remarkably, the highest similarities are found within the HighConfidenceSet. These results were confirmed for pairs with a maximum network distance ranging from one to four³⁵⁴.

When comparing the different interaction network types, we observed that the similarities of the pairs identified using the STRING network are lower than the similarities identified using BioMyn. A possible explanation is that functional associations as derived from gene co-expression data, which are contained in STRING but not in BioMyn, do not necessarily imply strong functional similarity. Nevertheless, the same overall tendency can be detected in results using the STRING network, that is, host factor pairs are more similar to each other than random pairs³⁵⁴.

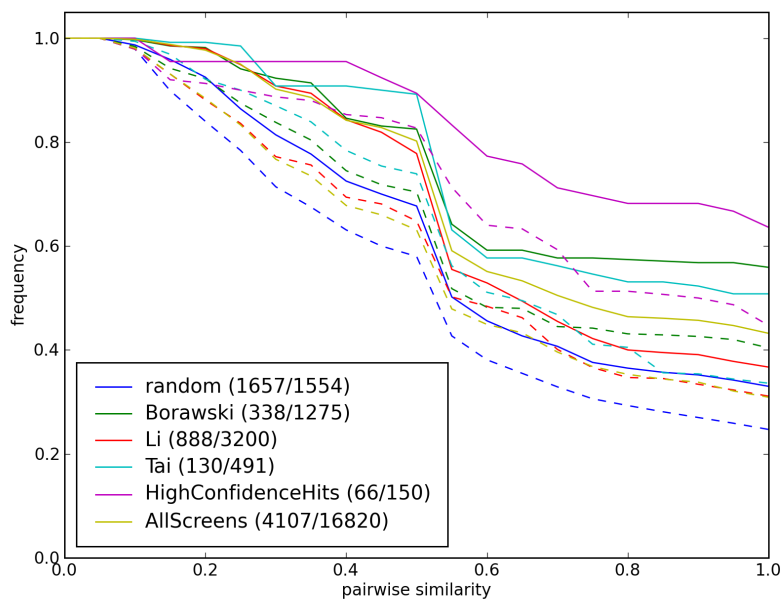


Figure 5.3: Normalized, cumulative functional similarity distributions of host factor pairs with a network distance of at most two³⁵⁴. Borawski⁴², Li²¹⁸, and³⁷⁰ denote host factors determined in the respective publications. The pairwise protein similarity is measured using sim_{Rel} based on GO molecular function annotations. Continuous lines display the pairs identified in the BioMyn network, dashed lines the pairs identified in the STRING network. Numbers in parentheses denote the number of identified pairs (BioMyn/STRING).

Conclusions Our functional network analyses revealed that pairs of published human host factors for viral infections are significantly more similar to each other than randomly selected protein pairs. This finding may be used for the validation and prioritization of experimentally determined sets of host factor candidates. Importantly, by integrating different types of biological data, namely, molecular interactions, known host factors, and functional protein annotations, the reliability of individual experimental results can be measured and improved to gain deeper insights into the molecular basis of viral infections in human host cells.

5.3 Application studies

This section describes the analysis of two large-scale screens determining host factors for the hepatitis C virus (HCV). In the following, the experiments described in Section 5.3.1 will be abbreviated as *kinome screen*³²¹, the recent analysis presented

in [Section 5.3.2](#) will be referred to as *druggable screen* (Poenisch et al., *in prep.**).

5.3.1 Human kinome screen for host factors required in HCV entry and replication

Our first application study was the analysis of an RNAi screen investigating the effect of human kinases on HCV infection³²¹. This *kinome screen* was performed in the group headed by Ralf Bartenschlager at the University of Heidelberg.

Experimental details The kinome screen focused on the early stages of the HCV life cycle, investigating the effect of host factors on viral entry or replication. In the primary screen, 719 known or predicted human kinases were targeted by three different small interfering RNAs (siRNAs) per gene. In addition to these 2157 siRNAs, positive and negative controls were added, for which a strong or no effect could be expected, respectively. The well-known entry host factor CD81²⁸ and two siRNAs that directly target HCV were used as positive controls, siRNA sequences without any complementary target sequence in the genome were used as negative controls.

Based on the results from twelve screen replicates, *z*-scores were calculated for each of the genes and hits were defined having a *z* - *score* < -1 or *z* - *score* > 1 and a *p* - *value* < 0.05. This modest cut-off score resulted in 83 host dependency factor (HDF) candidates and 95 potential host restriction factors (HRFs) that were further tested. In a validation screen, these 178 candidates were tested with four new siRNAs and a more robust experimental setup. A hit was defined as validated, if at least two out of four siRNAs had a *z* - *score* < -2.5 or *z* - *score* > 2.5. These criteria resulted in 13 HDFs that could be confirmed. In a further siRNA screen using HCV pseudoparticles (HCVpps), the role of the 13 HDFs in viral entry or replication was tested. No candidate was found to have an effect on HCV entry, proposing a role in viral replication³²¹.

Annotation of validated host factors The 13 validated HDFs of the kinome screen are listed in [Table 5.2](#). We annotated the candidates with several functional datasets. Of particular importance was the overlap with the in-house database of published host factors and protein interactions with previously known host factors and viral proteins. A more detailed version of [Table 5.2](#) can be found in the Supplemental Data of [Reiss et al.](#)³²¹.

Network analysis In order to place the validated HDFs in a cellular context and determine specific pathways, we performed an integrative network analysis, combining

*Poenisch, M., Metz, P., Blankenburg, H., Ruggieri, A., Lee, J.Y., Rupp, D., Rebhan, I., Diederich, K., Kaderali, L., Domingues, F.S., Albrecht, M., Lohmann, V., Erfle, H. and Bartenschlager, R. Identification of HNRNPK as regulator of hepatitis C virus particle production, *in prep.*

Table 5.2: Host dependency factors validated in the kinome screen³²¹. For each host factor, the table lists the average siRNA z-score in the validation screen, whether it is expressed in liver³⁶⁴, and if it was found in a previous host factor screen. In addition, interactions to viral proteins or previously reported host factors are reported.

Gene ID	Gene symbol	Score validation	Expressed in liver	Previously reported	PPI HCV ⁸⁴	PPI previous HDF	Co-complex previous HDF
1119	CHKA	-9.96	✓	218	-	-	-
5297	PI4KA	-7.82	✓	33,42,218,370,380,383	NS5A	-	✓
10000	AKT3	-5.47	✓	-	-	✓	✓
3101	HK3	-2.84	✓	-	-	-	-
1457	CSNK2A1	-4.551	✓	-	-	✓	✓
6725	SRMS	-4.15	-	-	-	-	-
9874	TLK1	-4.05	✓	-	-	-	✓
53944	CSNK1G1	-3.34	-	-	-	-	✓
5108	PCM1	-3.45	✓	-	-	-	✓
7075	TIE1	-3.23	✓	-	-	-	-
79705	LRRK1	-2.41	✓	-	-	-	✓
55750	AGK	-2.43	✓	-	-	✓	-
5214	PFKP	-2.86	✓	-	-	-	-

information about protein interactions and complexes with functional annotations. In particular, we incorporated all previously published HCV host factors that were available at the time (see Table 5.1). Eight of the 13 validated HDFs were found to be connected to at least one previously published host factor either by direct interaction or a co-complex membership (see Figure 5.4). We further investigated, if validated HDFs are known to be involved in specific cellular pathways. Two of the 13 HDFs were annotated to be involved in the ERBb pathway, which had been described as important for flaviviruses before²¹⁸.

For further analyses, we created a new host factor set by combining our confirmed HDFs with all previously reported HDFs. This allowed us to identify pathways and other functional annotations that are overrepresented within this set. A selection of the pathways found this way are shown in Figure 5.4, the full results can be found in the Supplemental Data of Reiss et al.³²¹.

PI4KIII α Choline kinase alpha (CHKA) and phosphatidylinositol 4-kinase III alpha (PI4KIII α), the two validated HDFs that showed the strongest effect in the kinome screen (see Figure 5.4 and Table 5.2), are both known to be involved in lipid membrane biosynthesis³²¹. While for CHKA no connection to other host factors could be established, PI4KIII α had interactions with several previously published host factors. Importantly, it had been reported in no less than six other RNAi screens for HCV host factors, which had all been published during the course of the study (see Table 5.2). However, despite this large body of evidence, the particular mechanism

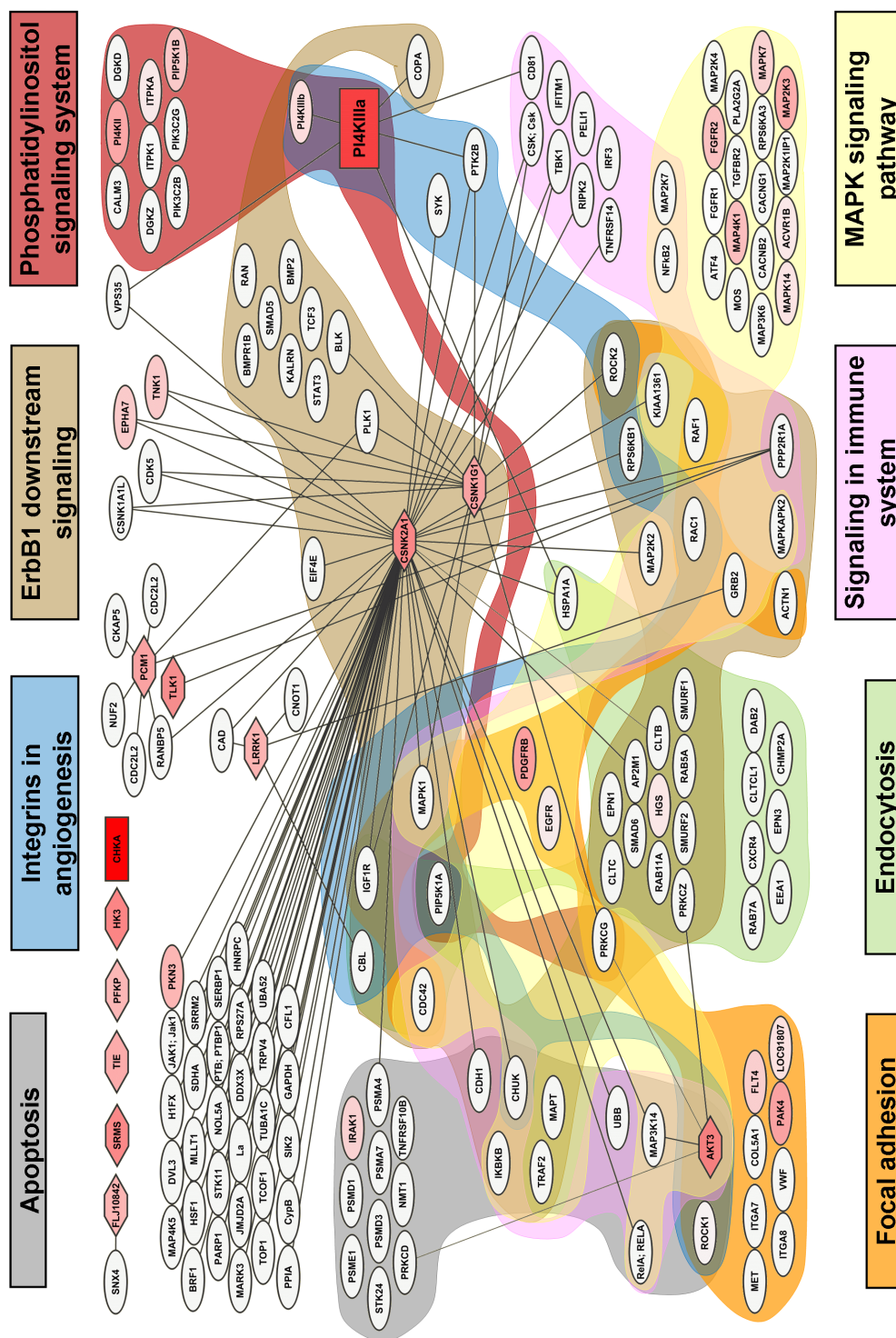


Figure 5.4: Network representation of selected hepatitis C virus (HCV) host factors³²¹. Host factors are depicted as nodes, connecting edges represent direct protein-protein interactions or co-membership in a protein complex. Node shapes symbolize if host factors were newly identified (octagonal), previously known (oval), or confirmed (rectangular). The strength of the inhibitory effect in the kinase screen is color-coded in the nodes (white: not tested; light red: weak effect; deep red: strong effect). A selection of the cellular pathways that are overrepresented among all HCV host factors are highlighted, along with the phosphatidylinositol 4-kinase III alpha (PI4KIII α) host factor.

by which PI4KIII α is important for HCV replication was not known.

Using a number of additional experiments, described in detail in Reiss et al.³²¹, our experimental collaboration partners could show that the enzymatic activity of PI4KIII α is required for HCV replication independently of the HCV genotype. Further, they showed that the absence of PI4KIII α leads to morphological changes in the membranous web structures, which are central for the establishment of HCV replication complexes. In particular, they found that HCV directly interacts with PI4KIII α , thereby stimulating its kinase activity. All this evidence combined suggest that HCV uses the kinase in order to change cellular structures in a way that favors its life cycle.

5.3.2 Human druggable-genome screen capturing all stages of the HCV life cycle

The second large-scale RNA interference screen that we supported with our analysis pipeline was also performed in the group headed by Ralf Bartenschlager at the University of Heidelberg. This screen used a different setup in order to test all stages of the viral life cycle. In particular, the set of genes that were tested for their potential role in HCV infection was much larger, consisting of more than 9 000 genes of the so-called *druggable genome*¹⁵⁹.

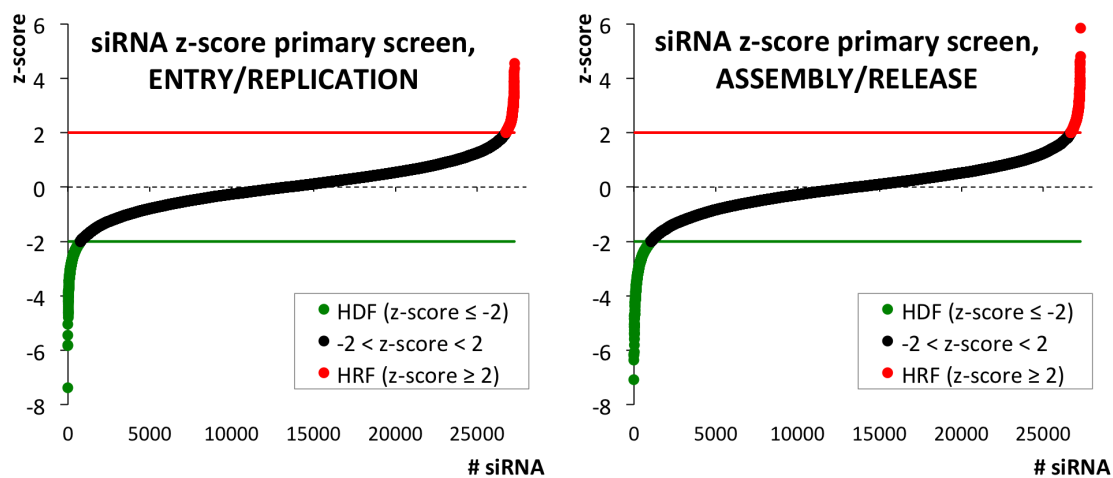


Figure 5.5: Score distribution of all siRNAs tested in the HCV druggable primary screen for entry or replication (left) and assembly or release (right). Hit siRNAs were defined by a z -score ≤ -2 (green line) for host dependency factors (HDF, green dots) or z -score ≥ 2 (red line) for host restriction factors (HRF, red dots) (Poenisch et al., *in prep.*).

Primary screen 9 102 genes of an extended druggable genome library were tested with three siRNAs per gene in the primary screen. The screen was composed of two parts, the first part investigating the early life cycle stages entry and replication, the second part focusing on the latter stages assembly and release of viral particles.

While the first part could be performed as described in the kinase screen (see [Section 5.3.1](#)), a reinfection of naïve cells with cultured supernatant was required to test the latter stages of the life cycle. Using a cut-off score of $z - score \leq -2$ and $z - score \geq 2$, 78 potential host dependency factors (HDFs) and 29 putative host restrictions factors (HRFs) were identified (see [Figure 5.5](#)).

We characterized the 106 host factor candidates of the primary screen by annotating them with GO terms and UniProt keywords (see [Figure 5.6](#)). The most common biological processes were transport, transcription and transcription regulation, whereas the most common molecular functions were hydrolase and transferase. For UniProt keywords, we did not filter for any particular category, but aimed to provide a broad overview of the cellular roles.

Meta-analysis The common procedure in RNAi screening is to investigate all hits from the primary screen, which have been selected using a certain cut-off score criteria, in a smaller validation screen²⁴⁶. Here, we wanted to complement this approach by selecting only a part of the primary hits and adding other genes for which there is a strong association to HCV. The rationale is that results will be biologically more relevant if they have been detected by different experimental protocols and procedures.

To this end, we conducted a computational meta-analysis that combined the results of the primary screen with our in-house database of published host factors, results of a protein-protein interaction study, comparative transcriptome profiles of HCV-infected and uninfected cells, and analyses of the components of the HCV-induced replication complexes (Poenisch et al., *in prep.*). We combined the individual data sources in a weighted summing approach, individual weight factors were assigned by the expert virologists of the Bartenschlager group. Considering a reasonable size of the validation screen, we selected 109 candidates from this meta-analysis for further validation.

Validation screen The 109 candidates from the meta-analysis were complemented with 108 hits from the primary screen. As nine genes were present in both lists, a total of 204 genes were tested in the validation screen. The screen was performed in a similar layout as the primary screen, but in a smaller, more reliable format. Four new siRNAs were selected, even changing the siRNA vendor to reduce potential biases. Candidates were defined as hits in the validation screen if two of their four siRNAs had a $z - score \leq -2$ for HDFs or $z - score \geq 2$ for HRFs (see [Figure 5.7](#)).

Considering that hits from the validation screen had to be consistent with their results from the primary screen or the meta-analysis, we confirmed 40 HDFs and 16 HRFs. As three of those were found to be able to act as both HDF and HRF, a total of 53 host factors were confirmed (see [Table 5.3](#)).

In order to gain insights into the cellular pathways involved, we annotated the confirmed host factors with functional datasets and computed enrichments (see [Ta-](#)

Table 5.3: Host factors confirmed in the HCV druggable screen. Each line summarizes the results from previous screens and our primary, meta, and validation screen. Abbreviations used: HDF = host dependency factor; HRF = host restriction factor; I = entry or replication, II = assembly or release; V = Validation screen, M = selected based on meta-analysis; D/d = HDF with 2 or 1 hit siRNAs; R/r = HRF with 2 or 1 hit siRNAs; P/p = selected as hit in primary screen with 2 or 1 siRNAs (Poenisch et al., *in prep.*).

		Primary screen				Validation screen						
Gene		I		II				I		II		
ID	Symbol	HDF	HRF	HDF	HRF	V-Setup	HDF	HRF	HDF	HRF	Confirmed	Previous
3190	HNRNPk	D	-	-	r	M,P	D	-	-	R	DR	HRF-II ²¹⁸
9793	CKAP5	d	-	D	-	M,P	-	R	D	-	D	HDF-I ³⁷⁰
60559	SPCS3	-	-	D	-	M,P	-	R	D	-	D	HDF-II ²¹⁸
9424	KCNK6	-	-	D	-	P	D	r	d	-	D	-
3679	ITGA7	-	-	d	-	M	d	-	D	-	D	HDF-I ^{42,218,370}
3851	KRT4	d	-	d	-	M	-	-	D	-	D	HDF-I ²¹⁸
4839	NOP2	D	-	d	-	P	-	-	D	-	D	HDF-I ²¹⁸
55361	PI4K2A	-	-	D	r	P	-	-	D	-	D	HDF-I ²²⁸
6236	RRAD	-	-	D	-	P	-	r	D	-	D	-
11142	PKIG	D	-	-	-	P	d	-	D	-	D	-
353274	ZNF445	D	-	-	-	P	d	-	D	-	D	-
23341	DNAJC16	-	-	D	-	M,P	-	-	D	-	D	HDF-I ⁴²
56992	KIF15	-	-	D	-	P	-	-	D	-	D	-
7554	ZNF8	-	-	D	-	P	-	-	D	-	D	-
7419	VDAC3	-	-	D	-	M,P	-	-	D	-	D	-
57463	AMIGO1	-	-	D	-	P	-	-	D	-	D	-
377677	CA13	-	-	D	-	P	-	-	D	-	D	-
84790	TUBA1C	-	-	-	-	M	-	R	D	-	DR	HDF-I ³⁷⁰
9588	PRDX6	-	-	-	-	M	-	R	D	-	DR	HDF-I ³⁸³
6233	RPS27A	-	-	-	-	M	-	r	D	-	D	HDF-I ⁴²
2130	EWSR1	-	-	-	-	M	d	r	D	-	D	HDF-I ²¹⁸
83694	RPS6KL1	-	-	-	-	M	d	-	D	-	D	HDF-I ^{42,218}
9475	ROCK2	-	-	-	-	M	-	-	D	-	D	HDF-I ^{33,74,218}
3303	HSPA1A	-	-	-	-	M	-	r	D	-	D	HDF-I ²¹⁸
10528	NOP56	-	-	-	-	M	d	-	D	-	D	HDF-I ²¹⁸
2885	GRB2	-	-	-	-	M	-	r	D	-	D	HDF-I ³¹⁴
9789	SPCS2	-	-	-	-	M	-	r	D	-	D	HDF-II ²¹⁸
9524	TECR	-	-	-	-	M	-	r	D	-	D	HDF-I ²¹⁸
7511	XPNPEP1	-	-	-	-	M	-	-	D	-	D	HDF-I ^{218,383}
2197	FAU	-	-	-	-	M	-	-	D	-	D	HDF-I ^{42,218}
1445	CSK	-	-	-	-	M	-	-	D	-	D	HDF-I ³⁶⁵
3881	KRT31	-	-	-	-	M	-	-	D	-	D	HDF-I ³⁷⁰
22907	DHX30	-	-	-	-	M	-	-	D	-	D	HDF-I ³⁸³
1147	CHUK	-	-	-	-	M	-	-	D	-	D	HDF-I ²¹⁸
10847	SRCAP	-	-	-	-	M	-	-	D	-	D	HDF-I ³¹⁴
65083	NOL6	-	-	-	-	M	-	-	D	-	D	HDF-I ²¹⁸
1381	CRABP1	d	-	d	-	p	D	-	D	-	D	-
84148	KAT8	d	-	-	R	P	D	-	-	-	D	-
348	APOE	-	r	d	-	p	-	r	D	-	D	-
55679	LIMS2	-	-	d	-	p	-	r	D	-	D	-
6428	SRSF3	-	-	-	R	P	-	r	-	R	R	-
9775	EIF4A3	-	-	-	R	P	-	R	-	r	R	-
7003	TEAD1	-	-	-	R	P	-	R	d	-	R	-
81788	NUAK2	-	-	-	-	M	-	R	d	r	R	HDF-I ^{218,262,370}
8775	NAPA	-	-	-	-	M	-	R	d	R	R	HDF-I ³⁷⁰
5901	RAN	d	-	-	-	M	-	R	-	-	R	HDF-I ³⁷⁰
7186	TRAF2	-	-	-	-	M	-	R	d	-	R	HDF-I ²⁶²
2107	ETF1	-	-	-	-	M	-	R	d	-	R	HDF-I ²¹⁸
1213	CLTC	-	-	-	-	M	-	R	d	-	R	HDF-I ³⁸⁰
2052	EPHX1	-	-	-	-	M	-	R	d	-	R	-
1314	COPA	-	-	-	-	M	-	R	-	-	R	HDF-I ³⁷⁰
29088	MRPL15	-	-	-	-	M	-	R	-	-	R	HDF-I ²¹⁸
2033	EP300	-	r	-	r	p	-	R	d	-	R	-

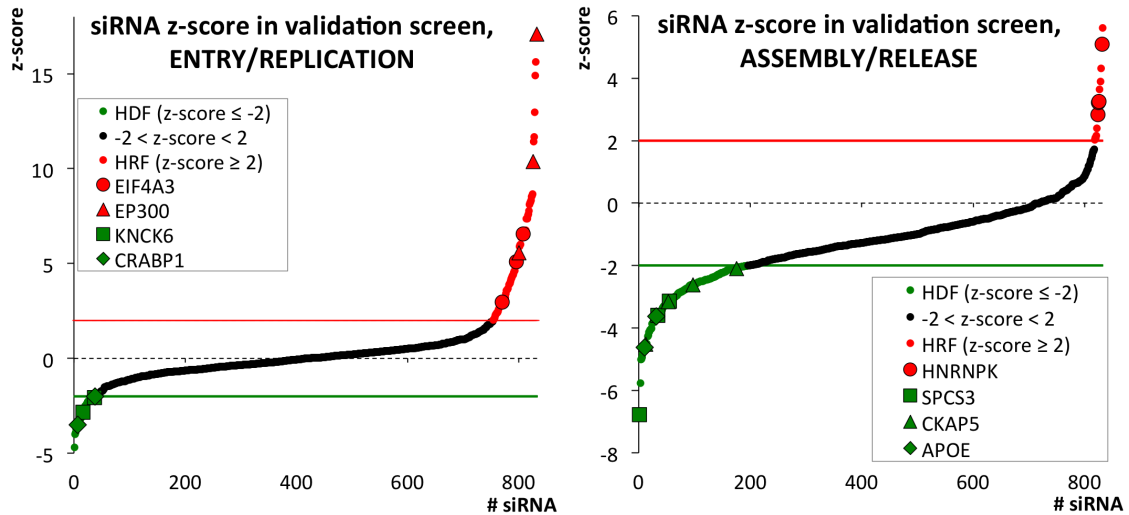


Figure 5.7: Distribution of all siRNAs tested in the druggable validation screen for entry or replication (left) and assembly or release (right). Hit siRNAs were defined by a z -score ≤ -2 (green line) in case of host dependency factors (HDF, green dots) or z -score ≥ 2 (red line) in case of host restriction factors (HRF, red dots). Genes targeted by siRNAs with very potent z -scores are highlighted (Poensch et al., *in prep.*).

ble 5.4). Amongst the biological processes that were annotated to the confirmed hits more frequently than expected was intracellular protein transport or the epidermal growth factor receptor signaling pathway. Both processes have already been associated with HCV before^{228,370}.

5.3.2.1 Further validation and role of HNRNPK

From the confirmed hits, ten candidates were selected for a more reliable validation setup. Candidates were selected if they were reported to have a known PPI with HCV or another previously reported host factor. This was complemented by

Table 5.4: Selection of biological processes enriched among validated hits in the druggable screen. Enrichments were computed with topGO, based on the total number of annotated (Annot.), significant (Signif.) and expected (Exp.) terms, using $LEA_{score} \leq 0.05$ as significance threshold⁴ (Poensch et al., *in prep.*).

GO ID	Term	Annot.	Signif.	Exp.	LEA
GO:0071702	organic substance transport	1969	18	6.68	4.30E-05
GO:0071840	cellular component organization or biogenesis	4290	28	14.55	6.20E-05
GO:0006886	intracellular protein transport	712	10	2.41	0.00011
GO:0034381	plasma lipoprotein particle clearance	31	3	0.11	0.00015
GO:0007173	epidermal growth factor receptor signaling pathway	160	5	0.54	0.0002
GO:0010988	regulation of low-density lipoprotein particle clearance	7	2	0.02	0.00023
GO:0033036	macromolecule localization	1873	16	6.35	0.0003
GO:0042059	negative regulation of epidermal growth factor receptor signaling pathway	39	3	0.13	0.00031
GO:0032092	positive regulation of protein binding	41	3	0.14	0.00036
GO:0044765	single-organism transport	2775	20	9.41	0.0004

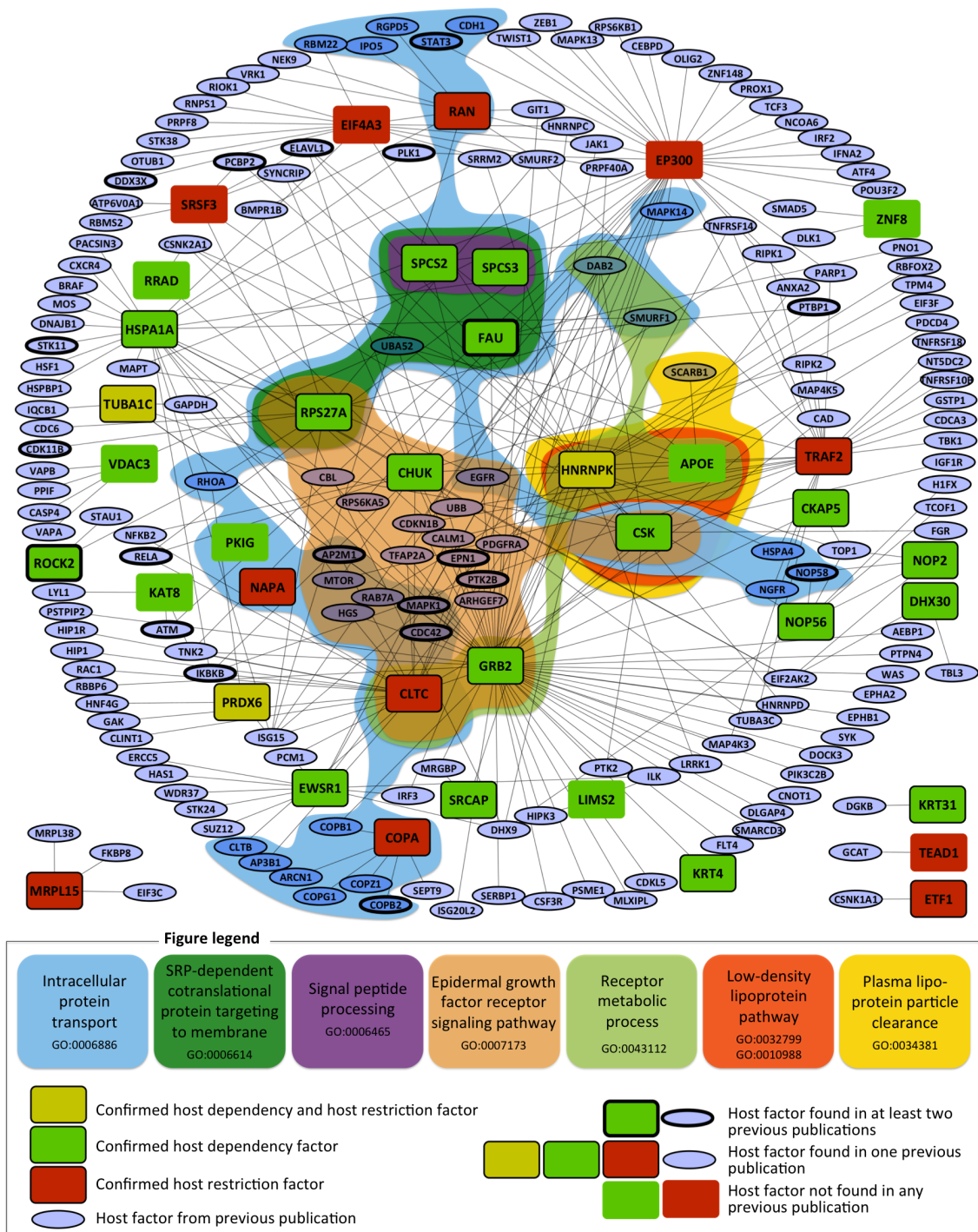


Figure 5.8: Integrated network visualization of confirmed host factors and enriched biological processes. Host factors found in the druggable screen (rectangular nodes) and in previous studies (oval or rectangular nodes with black borders) are shown. Oval nodes depict host factors that were either not investigated or not confirmed. Confirmed host dependency factors (HDF) are specified in green boxes, host restriction factors (HRF) in red boxes and factors with unclear function (HDF and/or HRF) in yellow boxes. Lines connecting two nodes indicate proteins with known physical interaction. Selected biological processes for which we found a significant enrichment of HCV host factors are highlighted by colored areas and listed with their Gene Ontology identifiers below the network in the correspondingly colored box (Poenisch et al., *in prep.*).

promising candidates, for which no role as host factor was reported before.

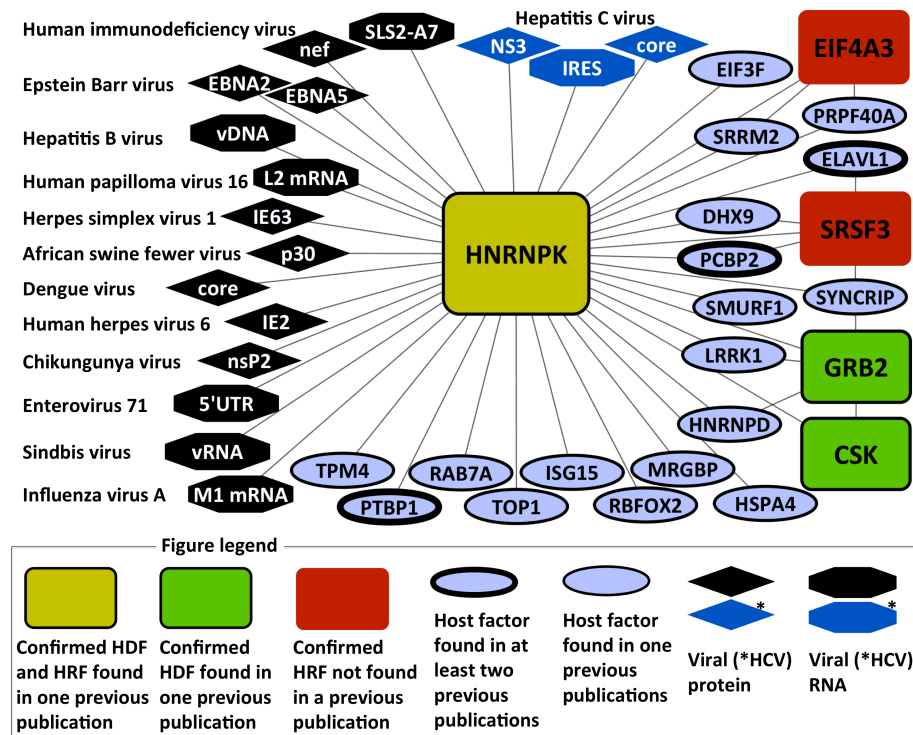


Figure 5.9: Interaction network of the heterogeneous nuclear ribonuclear protein K (HNRNPk), showing all human proteins (ovals and rectangles), viral proteins (diamonds) and viral RNAs (octagons) with reported interactions with HNRNPk. The depicted molecular interactions were extracted from the scientific literature by expert virologists (Poensch et al., *in prep.*).

The strongest effect was confirmed for the heterogeneous nuclear ribonuclear protein K (HNRNPk), which had already been reported in the primary screen and the meta-analysis. In the stringent validation screen, its role as host restriction factor (HRF) for viral assembly or release was confirmed. HNRNPk has already been associated with several different viral infections. For instance, it was reported to interact with the Sindbis virus⁵⁴ or the HCV proteins core¹⁶² and NS3⁸⁴. Information about previously reported interactions with human or viral proteins and nucleic acids were extracted from the scientific literature by our collaboration partners; the resulting network is shown in Figure 5.9.

Given the consistent effect of HNRNPk during all stages of the druggable screen and the reported evidence for a role in viral infections, our experimental collaboration partners in Heidelberg went on to further characterize this host factor. First, they investigated any potential effects on viral entry or replication and made sure that the restrictive effect of HNRNPk is independent of the HCV genotype. Unlike in other viral infections as caused by the aforementioned Sindbis virus⁵⁴, they did not find a change in HNRNPk's subcellular localization after infection. In additional characterization studies, they found that the two protein domains of the

modular HNRNPK are particularly required for the restriction of HCV and that the protein interacts with viral RNA. While the exact mechanism is still not fully understood, these results suggest that HNRNPK regulates the production of viral particles in the assembly process by controlling the availability of viral RNA.

5.4 Conclusions

5.4.1 Summary

Viruses and the infections they cause are a major threat to human health. While intense research efforts have resulted in improved treatment options²⁰⁵, no vaccines are yet available for important viruses like those causing the acquired immune deficiency syndrome or hepatitis C^{24,241}. Classic antiviral drugs commonly operate by activating the immune response, which has limited desired and strong side-effects, or by targeting parts of the viral genome or proteome²³³. The pressure of direct antiviral drugs leads to the emergence of resistance mutations, which due to the poor viral error-correction in the replication machinery and the rapid turnover may quickly lead to viral populations that are drug resistant⁴⁰⁴.

Due to their compact genomes, viruses rely on the human cellular machinery to fulfill their life cycle²⁴⁵. Consequently, human host factors that are required for viral infections have emerged as a promising research target. First, knowledge about these factors will advance our basic understanding of how the virus uses and abuses the cell¹¹³. Second, host factors may provide a new treatment option for antiviral therapy¹⁹⁵. While targeting those host factors may have undesired side-effects for the normal function of human cells, the virus cannot simply evade this pressure by developing resistance mutations, but has to completely adapt its life cycle to the new situation⁹⁷.

RNA interference (RNAi) is the detection method that has been used in most experimental screens for determining host factors for different viruses, including Dengue, West Nile, Human Immunodeficiency, Hepatitis C and Influenza⁷⁰. RNAi is a powerful technique that allows investigating the impact of individual genes on the viral life cycle by temporarily degrading their gene products. However, RNAi studies are associated with significant false positive and false negative results, mostly due to undesired off-target effects¹⁴⁶. This has resulted in a remarkably low overlap of studies investigating the same viral systems⁵⁵. As a consequence, in addition to a robust experimental setup that includes a sufficient number of replicates and independent siRNAs per gene, statistical post-processing and bioinformatics analyses are required to ensure that results have a biological relevance.

To this end, we have developed an integrated analysis pipeline that combines different data types and analysis approaches. Our in-house database containing multiple previously published host factor screens turned out to be of particular importance, enabling us to place individual screen results into a broader context. In

a small study, we investigated the functional annotations of host factors contained in this database and found that the pairwise functional similarity of host factors is significantly higher than for randomly selected pairs³⁵⁴. A high confidence subset of host factors, which had been found in multiple independent studies, turned out to have the highest similarities, proposing such an approach as means for validating the results of individual studies.

We further applied our analysis pipeline to two large-scale screens for host factors of the hepatitis C virus. In both of these screens, we supported the whole analysis process and prioritized the most-promising candidates for further characterization studies. In the more recent druggable screen (Poenisch et al., *in prep.*), which investigated the complete viral life cycle, we further influenced the setup of the validation screen by performing a meta-analysis that contributed additional host factor candidates that had not been investigated in the primary screen.

Both studies yielded several promising host factor candidates. In particular, Reiss et al. found that HCV interacts with the lipid kinase phosphatidylinositol 4-kinase III alpha (PI4KIII α) that was consistently found to have strong effects in different RNAi screens, thereby stimulating its kinase activity³²¹. As the absence of PI4KIII α resulted in changed morphology of the viral replication complexes, a role of the kinase in the formation of these complexes is likely. Poenisch et al. recently determined the role of the heterogeneous nuclear ribonucleoprotein K (HNRNPK) as a restricting factor in the assembly of viral particles (Poenisch et al., *in prep.*). While the exact mechanism is not fully understood, the experiments propose a role in regulating the availability of HCV RNA in the production of viral particles.

5.4.2 Outlook

Despite greatly improved antiviral treatment options, human host factors for viral infections will remain an important research target. Most of the host factor screens in the past have been conducted in a classic approach, where a large primary screen was used to propose candidates for a more stringent validation screen, and few validated targets were characterized in detail. While this procedure proved its usefulness in detecting factors of biological relevance³²¹, it might miss candidates with weaker effects. For example, host factors that are slightly above or below a z-score cut-off in the primary screen are not further investigated, even though they might be of central importance to the virus. Complementing the results of RNAi screens with additional data types, as we have exemplarily shown with our meta-analysis approach, may help to uncover such cases.

Integrative analysis approaches may also be used to reduce the importance of selecting appropriate cut-off scores. Currently, the scores for separating hits from non-hits are selected more or less arbitrarily, often simply reusing values from previously published screens. Instead, functional data could be incorporated to flexibly select cut-off scores in a more objective fashion, for instance, based on changing functional similarities of host factors.

Data about specific siRNA sequences are currently not incorporated into analysis pipelines, although this information could be used to better detect unwanted off-target effects or to correlate resulting scores with different chemical siRNA properties³⁵⁰. [Buehler et al.](#) even proposed a method for detecting pathway components purely based on off-target effects⁵³.

A general problem of loss-of-function studies investigating the effect of single-gene knockouts are potential compensatory effects. Parallel paths in cellular pathways or genetic redundancy may re-establish the function that was lost as a consequence of the mRNA degradation. Experimental setups that allow testing two genes in parallel¹⁶¹ and improved analysis methods that incorporate detailed pathway maps might be able to improve results in this aspect²⁴⁶.

Further advancement in the integrated network analyses of host factor screens can be expected from the incorporation of additional data types. Of particular importance in this regard is high-quality data about drugs and their human targets, as this would promote finding host factor candidates that can easily be further experimentally characterized⁸³. In addition, information about disease associations might be used to complement the functional datasets that are currently used. Instead of identifying network clusters of host factors that are functionally similar, this would allow detecting modules of closely connected host factors for which there is a known disease association.

Our visualizations combining viral host factor networks with functional enrichments, presented in [Figures 5.8 and 5.4](#), were very well received by other bioinformaticians, asking for support in creating similar visualizations. As much of the work currently still involves laborious manual layout adjustments, a Cytoscape plugin that takes over large parts of the layout process would be a great help for many researchers.

6 Conclusions

This last chapter concludes the thesis by summarizing our methodological developments and the application studies that have been carried out. In addition, it points out particular aspects of interactomics, where further advancements may be made.

6.1 Summarizing remarks

This thesis described novel methods for biomedical data integration and their applications in human systems biology. This interdisciplinary research area utilizes the combined analysis of different molecular, biological, and medical data types to discover information that may remain hidden in individual analyses. The data type that was of central importance to our work was information about the interaction of proteins with other molecules. As molecular interactions are involved in almost every aspect of cellular functioning, knowledge of the complete human interactome, the set of all molecular interactions inside a cell, is crucial for understanding complex biological systems. Consequently, intense research has resulted in great amounts of interaction data that have been described in publications or stored in dedicated databases. The two main problems related to these datasets, which limit their value for research and applications, are their general availability and the varying quality. Within this thesis, we developed new methods that addressed both of these issues.

Chapter 3 described two systems that were developed in international collaborations to improve the availability of interaction data. Our Distributed System for Molecular Interactions (DASMI) is built on the concept of decentralization. Interaction data are stored in various online repositories and is retrieved, combined, and visualized upon user request^{40,41}. Compared to static integrative databases like iRefIndex or STRING, this approach ensures that interaction data can be easily maintained and kept up-to-date. In addition, interested interaction data providers may use our software to quickly set up additional servers for making their data available¹⁸¹. DASMI, developed in the context of the European BioSapiens Network of Excellence¹²³, was the successful prototype that provided the basis for a follow-up project of the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI). Today, this successor system named PSI Common Query Interface (PSICQUIC¹⁴) has a strong support in the interaction community and is

used in many projects. This means that the two distributed systems presented in this chapter profoundly improved the situation of interaction data availability.

Methodological advancements in the assessment of interaction data quality were described in [Chapter 4](#). The varying reliability of different interaction determination or prediction techniques necessitated independent means to qualitatively assess the resulting data. The great number and the diversity of approaches that had been proposed to this end, renders a comprehensive central scoring solution unfeasible. Consequently, we applied another decentralization strategy, in which scoring servers can be made available at remote locations. The servers are accessed from a central client, where users can upload individual interaction data that are to be scored and then retrieve the results. Building on experiences gained with a proof-of-concept scoring solution based on the aforementioned DASMI^{40,41}, we developed the PSI Common Confidence Scoring System (PSISCORE) in cooperation with the HUPO-PSI³⁸. Our open-source reference implementations facilitate the incorporation of new scoring algorithms for assessing different aspects of molecular interactions. The usefulness of our system was proven by the PSISCORE servers that have been made available in Italy, Norway, and the United Kingdom¹⁴.

[Chapter 5](#) demonstrated with several application studies how interactions and other types of biomedical data may be jointly analyzed to gain new insights of direct biological relevance. In particular, the chapter described the work performed in the analysis, prioritization, and interpretation of data produced by two large-scale RNA interference screens for determining human host factors for the hepatitis C virus (HCV), an infection affecting 180 million people worldwide³⁹. These studies proved that interdisciplinary collaborations between bioinformaticians and experimental research groups are increasingly important in our age of heterogeneous, systems-level datasets¹⁰⁵. Notably, we supported the discovery and functional characterization of novel host factors for HCV. This expanded our knowledge of the basic viral life cycle and provided promising new candidates for therapeutic advancements in antiviral treatments³²¹.

As the focus of this thesis was on the development of new methods for biomedical data integration, only a selection of the performed application studies were described. Another study together with research groups in Lübeck, Germany, and Bolzano, Italy, involved an integrative network analysis approach incorporating protein interactions and functional gene similarity networks for the prioritization of potential interaction partners of Parkin, a key protein in Parkinson's disease⁴²³. Furthermore, a fruitful collaboration with structural biologists in Budapest, Hungary, determined and prioritized new putative binding partners of mitogen-activated protein kinases (MAPKs) and extracellular signal-regulated kinases (ERKs), which form part of the important cellular MAPK/ERK pathway that transmits extracellular signals to the nucleus¹²⁷.

In summary, we developed a number of computational methods that successfully improved the availability and quality of interaction data and demonstrated how

novel biomedical data integration techniques may be applied to yield new insights of direct biological relevance in interdisciplinary collaborations.

6.2 Perspectives

Information about molecular interactions will remain central to our understanding of the human cell. In particular, it is still a long way towards a complete human interactome³⁸⁸, as current methods do not adequately reflect its time- and space-dependent nature, for example, not capturing that two proteins might only interact in particular cellular organelles after being co-expressed in response to a certain cellular state. Methodological advancements to detect and incorporate these data into existing frameworks will be the next challenge in interactomics³⁰⁵. This also includes the development of new visualization techniques that complement the static two-dimensional network representations.

The distributed systems for interaction data retrieval, PSICQUIC in particular, are increasingly being used in large-scale application studies. A further step into this direction would be made by software solutions that integrate results from PSICQUIC servers into local data warehouses. This would combine the benefits of decentralization, that is, recent data and the ability to incorporate new servers, with the superior analysis capabilities of data warehouses⁵⁸. Incorporating PSISCOPE into such an integration pipeline in order to assess the combined interaction data with different scoring algorithms would make an additional improvement. For certain computationally expensive algorithms, like those based on structural modeling, such a large-scale application might currently not be feasible.

From the two distributed HUPO-PSI systems dealing with molecular interactions, PSICQUIC has the stronger community support. While this is largely due to the fact that there are far more providers for interaction data than for scoring routines, another contributing factor is that PSICQUIC received greater investments in terms of personnel. We have learned over the years that, until a distributed system has reached a critical size and importance, a lot of time-consuming support is required to convince potential service providers to invest the additional effort that is required to set up an own server. Thus, by investing more manpower in PSISCOPE, its support in the interaction community is likely to grow, resulting in an improved selection of scoring servers. In addition, new clients for popular bioinformatics platforms like Cytoscape, Galaxy, or Bioconductor would allow the incorporation of confidence scoring in many existing workflows. The long-term goal of interaction data quality assessment, however, is the combination of various confidence scoring approaches into a meaningful combined score. While the outcome of this endeavor is uncertain, it will be a challenging research area over the next years.

In our different application studies, we have demonstrated how integrative analysis approaches yield new insights and aid researchers in answering particular biological questions. Further means for combining heterogeneous datasets such as statistical

learning methods might further improve the results. However, this improved performance is often accompanied by sacrificed interpretability, which can ultimately result in a black box that is no longer accepted by experimental cooperation partners. Generally, improved results in integrative analysis studies can be expected by incorporating additional data types, for example, information about drugs and drug targets or disease associations. For sufficiently small hit lists, like the top hits from our ranked HCV host factors screen results, structural models may be computationally predicted to elucidate putative interactions between host factor and viral proteins.

In summary, it can be expected that biomedical data integration will remain an important research discipline in the future, as new application areas are opened up by steadily evolving experimental techniques. For example, single-cell perturbations, combining sequencing, imaging, and proteomics experiments, will facilitate new large-scale system biological studies⁹⁹. Another area, which holds great promises, is the combination of clinical, biological, and molecular data, as it might be achieved using extended electronic patient records. Such a system would enable integrating data over multiple different scales, from free-text diagnoses and prescriptions to images created by X-ray or magnetic resonance imaging, from diagnostic blood parameters to individual patient sequence data and information about disease-associated single-nucleotide polymorphisms¹⁸². However, before this can become reality, a multitude of issues in the areas of data access, standards, and protocols need to be tackled, along with legal and ethical questions. All in all, interdisciplinary approaches for combining, processing, analyzing, and prioritizing such complex and heterogeneous datasets will remain crucial in the foreseeable future.

Bibliography

- [1] Agaisse, H., Burrack, L. S., Philips, J. A., Rubin, E. J., Perrimon, N., and Higgins, D. E. (2005). Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science*, 309(5738):1248–1251.
- [2] Akiva, E., Friedlander, G., Itzhaki, Z., and Margalit, H. (2012). A dynamic view of domain-motif interactions. *PLoS Computational Biology*, 8(1):e1002341.
- [3] Albrecht, M., Huthmacher, C., Tosatto, S. C. E., and Lengauer, T. (2005). Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21 Suppl 2:ii220–221.
- [4] Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607.
- [5] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutillier, K., Burgess, E., Buzadzija, K., Cavero, R., D’Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. F., and Hogue, C. W. V. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue):D418–D424.
- [6] Aloy, P., Ceulemans, H., Stark, A., and Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5):989–998.
- [7] Aloy, P. and Russell, R. B. (2002). Interrogating protein interaction networks

- through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5896–5901.
- [8] Aloy, P. and Russell, R. B. (2002). The third dimension for protein interactions and complexes. *Trends in Biochemical Sciences*, 27(12):633–638.
- [9] Aloy, P. and Russell, R. B. (2005). Structure-based systems biology: a zoom lens for the cell. *FEBS Letters*, 579(8):1854–1858.
- [10] Alvisi, G., Madan, V., and Bartenschlager, R. (2011). Hepatitis C virus and host cell lipids: An intimate connection. *RNA Biology*, 8(2):258–269.
- [11] Amberkar, S., Kiani, N. A., Bartenschlager, R., Alvisi, G., and Kaderali, L. (2013). High-throughput RNA interference screens integrative analysis: Towards a comprehensive understanding of the virus-host interplay. *World Journal of Virology*, 2(2):18–31.
- [12] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O’Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–D159.
- [13] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(Database issue):D525–D531.
- [14] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E. W., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O’Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., and Hermjakob, H. (2011). Psicquic and psiscore: accessing and scoring molecular interactions. *Nature Methods*, 8(7):528–529.
- [15] Ashburner, M., Ball, C., Blake, J., and Botstein, D. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.
- [16] Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

-
- [17] Bader, G. and Hogue, C. W. V. (2000). BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, 16(5):465–477.
- [18] Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(Database issue):D504–D506.
- [19] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 29(1):242–245.
- [20] Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85.
- [21] Bader, S., Kühner, S., and Gavin, A.-C. (2008). Interaction networks for systems biology. *FEBS Letters*, 582(8):1220–1224.
- [22] Barabási, A.-L. (2005). Sociology. Network theory—the emergence of the creative enterprise. *Science*, 308(5722):639–641.
- [23] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- [24] Barnes, E., Folgori, A., Capone, S., Swadling, L., Aston, S., Kurioka, A., Meyer, J., Huddart, R., Smith, K., Townsend, R., Brown, A., Antrobus, R., Ammendola, V., Naddeo, M., O’Hara, G., Willberg, C., Harrison, A., Grazioli, F., Esposito, M. L., Siani, L., Traboni, C., Oo, Y. Y., Adams, D., Hill, A., Colloca, S., Nicosia, A., Cortese, R., and Klenerman, P. (2012). Novel adenovirus-based vaccines induce broad and sustained T cell responses to HCV in man. *Science Translational Medicine*, 4(115):115ra1.
- [25] Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, 307(5715):1621–1625.
- [26] Bartenschlager, R. (2006). Hepatitis C virus molecular clones: from cDNA to infectious virus particles in cell culture. *Current Opinion in Microbiology*, 9(4):416–422.
- [27] Bartenschlager, R., Penin, F., Lohmann, V., André, P., Andre, P., and Lyon, D. (2010). Assembly of infectious hepatitis C virus particles. *Trends in Microbiology*, 19(2):95–103.

- [28] Bartosch, B., Vitelli, A., Granier, C., Goujon, C., Dubuisson, J., Pascale, S., Scarselli, E., Cortese, R., Nicosia, A., and Cosset, F.-L. (2003). Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor. *The Journal of Biological Chemistry*, 278(43):41624–41630.
- [29] Bassendine, M. F., Sheridan, D. a., Felmlee, D. J., Bridge, S. H., Toms, G. L., and Neely, R. D. G. (2011). HCV and the hepatic lipid pathway as a potential treatment target. *Journal of Hepatology*, 55(6):1428–1440.
- [30] Belda, O. and Targett-Adams, P. (2012). Small molecule inhibitors of the hepatitis C virus-encoded NS5A protein. *Virus Research*, 170(1-2):1–14.
- [31] Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:i38–46.
- [32] Ben-Hur, A. and Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2.
- [33] Berger, K. L., Cooper, J. D., Heaton, N. S., Yoon, R., Oakland, T. E., Jordan, T. X., Mateu, G., Grakoui, A., Randall, G., and Consulting, S. (2009). Roles for endocytic trafficking and phosphatidylinositol 4-kinase III alpha in hepatitis C virus replication. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7577–82.
- [34] Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2013). The future of the protein data bank. *Biopolymers*, 99(3):218–222.
- [35] Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366.
- [36] Birmingham, A., Anderson, E. E. M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., Marshall, W. S., and Khvorova, A. (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods*, 3(3):199–204.
- [37] Blankenburg, H. (2007). A distributed annotation system for molecular interactions. Diploma thesis, Universität Heidelberg; Hochschule Heilbronn.
- [38] Blankenburg, H. and Albrecht, M. (2013). PSIScore (Quality Scoring of Protein Interactions). In Dubitzky, W., Wolkenhauer, O., Cho, K.-H., and Yokota, H., editors, *Encyclopedia of Systems Biology*, pages 1801–1802. Springer New York.

-
- [39] Blankenburg, H., Diehl, S., Ramírez, F., Wörz, I., Poenisch, M., Bartenschlager, R., and Albrecht, M. (2010). Discovery and prioritization of human cellular factors required for HCV infection. *New Biotechnology*, 27 Suppl 1:S24–S25.
- [40] Blankenburg, H., Finn, R. D., Prlić, A., Jenkinson, A. M., Ramírez, F., Emig, D., Schelhorn, S.-E., Büch, J., Lengauer, T., and Albrecht, M. (2009). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328.
- [41] Blankenburg, H., Ramírez, F., Büch, J., and Albrecht, M. (2009). DASMIweb: online integration, analysis and assessment of distributed protein interaction data. *Nucleic Acids Research*, 37(Web Server issue):W122–W128.
- [42] Borawski, J., Troke, P., Puyang, X., Gibaja, V., Zhao, S., Mickanin, C., Leighton-Davies, J., Wilson, C. J., Myer, V., Cornellataracido, I., Baryza, J., Tallarico, J., Joberty, G., Bantscheff, M., Schirle, M., Bouwmeester, T., Mathy, J. E., Lin, K., Compton, T., Labow, M., Wiedmann, B., and Gaither, L. A. (2009). Class III phosphatidylinositol 4-kinase alpha and beta are novel host factor regulators of hepatitis C virus replication. *Journal of Virology*, 83(19):10058–10074.
- [43] Bordner, A. J. and Gorin, A. A. (2008). Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. *BMC Bioinformatics*, 9:234.
- [44] Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Letters*, 286(I):47–54.
- [45] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–9.
- [46] Börner, K., Hermle, J., Sommer, C., Brown, N. P., Knapp, B., Glass, B., Kunkel, J., Torralba, G., Reymann, J., Beil, N., Beneke, J., Pepperkok, R., Schneider, R., Ludwig, T., Hausmann, M., Hamprecht, F., Erfle, H., Kaderali, L., Kräusslich, H.-G., and Lehmann, M. J. (2010). From experimental setup to bioinformatics: an RNAi screening platform to identify host factors involved in HIV-1 replication. *Biotechnology Journal*, 5(1):39–49.
- [47] Brass, V., Moradpour, D., and Blum, H. E. (2006). Molecular virology of hepatitis C virus (HCV): 2006 update. *International Journal of Medical Sciences*, 3(2):29–34.
- [48] Braun, P. (2012). Interactome mapping for analysis of complex phenotypes: Insights from benchmarking binary interaction assays. *Proteomics*, 12(10):1499–1518.

- [49] Braun, P., Tasan, M., and Dreze, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods*, 6(1):91–97.
- [50] Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–82.
- [51] Brown, K. R. and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8(5):R95.
- [52] Browne, F., Wang, H., Zheng, H., and Azuaje, F. (2009). GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction. *Source Code for Biology and Medicine*, 4:2.
- [53] Buehler, E., Khan, A. A., Marine, S., Rajaram, M., Bahl, A., Burchard, J., and Ferrer, M. (2012). siRNA off-target effects in genome-wide screens identify signaling pathway members. *Scientific Reports*, 2:428.
- [54] Burnham, A. J., Gong, L., and Hardy, R. W. (2007). Heterogeneous nuclear ribonuclear protein K interacts with Sindbis virus nonstructural proteins and viral subgenomic mRNA. *Virology*, 367(1):212–221.
- [55] Bushman, F. D., Malani, N., Fernandes, J., D’Orso, I., Cagney, G., Diamond, T. L., Zhou, H., Hazuda, D. J., Espeseth, A. S., König, R., Bandyopadhyay, S., Ideker, T., Goff, S. P., Krogan, N. J., Frankel, A. D., Young, J. A. T., and Chanda, S. K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathogens*, 5(5):e1000437.
- [56] Butland, G., Li, J., and Yang, W. (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, 433(7025):531–537.
- [57] Buza, T. J., McCarthy, F. M., Wang, N., Bridges, S. M., and Burgess, S. C. (2008). Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Research*, 36(2):e12.
- [58] Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nature Methods*, 10(8):690–691.
- [59] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266.
- [60] Casado-Vela, J., Matthiesen, R., Sellés, S., and Naranjo, J. (2013). Protein-protein interactions: gene acronym redundancies and current limitations precluding automated data integration. *Proteomes*, 1(1):3–24.

-
- [61] Ceol, A., Chatr-Aryamontri, A., Licata, L., and Cesareni, G. (2008). Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Letters*, 582(8):1171–1177.
- [62] Chang, D. T.-H., Syu, Y.-T., and Lin, P.-C. (2010). Predicting the protein-protein interactions using primary structures with predicted protein surface. *BMC Bioinformatics*, 11 Suppl 1:S3.
- [63] Chao, T.-C., Su, W.-C., Huang, J.-Y., Chen, Y.-C., Jeng, K.-S., Wang, H.-D., and Lai, M. M. C. (2012). Proline-serine-threonine phosphatase-interacting protein 2 (PSTPIP2), a host membrane-deforming protein, is critical for membranous web formation in hepatitis C virus replication. *Journal of Virology*, 86(3):1739–1749.
- [64] Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O’Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue):D816–D823.
- [65] Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(Database issue):D572–D574.
- [66] Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E. E., Futschik, M. E., and Ha, C. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Research*, 35(Database issue):D590–D594.
- [67] Chaurasia, G., Malhotra, S., Russ, J., Schnoegl, S., Hänig, C., Wanker, E. E., and Futschik, M. E. (2009). UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. *Nucleic Acids Research*, 37(Database issue):D657–D660.
- [68] Chen, N., Harris, T. W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.-K., Chen, W. J., Cunningham, F., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H.-M., Nakamura, C., Pai, S., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E. M., Van Auken, K., Wang, Q., Durbin, R., Spieth, J., Sternberg, P. W., and Stein, L. D. (2005). WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Research*, 33(Database issue):D383–D389.
- [69] Chen, Y.-C., Su, W.-C., Huang, J.-Y., Chao, T.-C., Jeng, K.-S., Machida, K., and Lai, M. M. C. (2010). Polo-like kinase 1 is involved in hepatitis C virus replication by hyperphosphorylating NS5A. *Journal of Virology*, 84(16):7983–7993.

- [70] Cherry, S. (2009). What have RNAi screens taught us about viral-host interactions? *Current Opinion in Microbiology*, 12(4):446–452.
- [71] Choo, Q. L., Kuo, G., Weiner, a. J., Overby, L. R., Bradley, D. W., and Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, 244(4902):359–362.
- [72] Chua, H. N. and Wong, L. (2008). Increasing the reliability of protein interactomes. *Drug Discovery Today*, 13(15-16):652–658.
- [73] Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, 20(3):426–427.
- [74] Coller, K. E., Berger, K. L., Heaton, N. S., Cooper, J. D., Yoon, R., and Randall, G. (2009). RNA interference and single particle tracking analysis of hepatitis C virus endocytosis. *PLoS Pathogens*, 5(12):e1000702.
- [75] Coller, K. E., Heaton, N. S., Berger, K. L., Cooper, J. D., Saunders, J. L., and Randall, G. (2012). Molecular determinants and dynamics of hepatitis C virus secretion. *PLoS Pathogens*, 8(1):e1002466.
- [76] Côté, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, 8:401.
- [77] Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., Biankin, A. V., Hautaniemi, S., and Wu, J. (2012). PINA v2.0: mining interactome modules. *Nucleic Acids Research*, 40(Database issue):D862–D865.
- [78] Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*, 14 Spec No(2):R171–R181.
- [79] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhoute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. (2009). Addendum: Literature-curated protein interaction datasets. *Nature Methods*, 6(12):934–935.
- [80] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhoute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. (2009). Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46.

-
- [81] Dahari, H., Sainz, B., Perelson, A. S., and Uprichard, S. L. (2009). Modeling subgenomic hepatitis C virus RNA kinetics during treatment with alpha interferon. *Journal of Virology*, 83(13):6383–6390.
- [82] Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92.
- [83] de Chasseay, B., Meyniel-Schicklin, L., Aublin-Gex, A., André, P., and Lotteau, V. (2012). New horizons for antiviral drug discovery from virus-host protein interaction networks. *Current Opinion in Virology*, 2(5):606–613.
- [84] de Chasseay, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugué, S., Meiffren, G., Pradezynski, F., Faria, B. F., Chantier, T., Le Breton, M., Pellet, J., Davoust, N., Mangeot, P. E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P. O., Vidal, M., André, P., Rabourdin-Combe, C., and Lotteau, V. (2008). Hepatitis C virus infection protein network. *Molecular Systems Biology*, 4:230.
- [85] De Clercq, E. (2012). Human viral diseases: what is next for antiviral drug discovery? *Current Opinion in Virology*, 2(5):572–579.
- [86] Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, 1(5):349–356.
- [87] Debes, J. D. and Smith, C. I. (2012). NS5A: a new target for antiviral drugs in the treatment of hepatitis C virus infection. *Hepatology*, 56(3):797–799.
- [88] Del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R. C., Galeota, E., Lounay, G., Goll, J., Breuer, K., Ono, K., Salwinski, L., and Hermjakob, H. (2013). A new reference implementation of the PSICQUIC web service. *Nucleic Acids Research*, 41(Web Server issue):W601–W606.
- [89] Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A. M., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J. S., Hope, I. A., Tissenbaum, H. A., Mango, S. E., and Walhout, A. J. M. (2006). A gene-centered *C. elegans* protein-DNA interaction network. *Cell*, 125(6):1193–1205.
- [90] Diamond, D. L., Syder, A. J., Jacobs, J. M., Sorensen, C. M., Walters, K.-A., Proll, S. C., McDermott, J. E., Gritsenko, M. a., Zhang, Q., Zhao, R., Metz, T. O., Camp, D. G., Waters, K. M., Smith, R. D., Rice, C. M., and Katze, M. G. (2010). Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics. *PLoS Pathogens*, 6(1):e1000719.
- [91] Dieterich, D. (2012). The end of the beginning for hepatitis C treatment. *Hepatology*, 207127(5):1–3.

- [92] Donaldson, I. M. (2010). Protein Interaction Data Resources. In Bradshaw, Ralph A. and Dennis, E. A., editor, *Handbook of Cell Signaling*, chapter 170, pages 1375–1385. Elsevier, 2 edition.
- [93] Doncheva, N. T., Kacprowski, T., and Albrecht, M. (2012). Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):429–442.
- [94] Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., and Stein, L. (2001). The Distributed Annotation System. *BMC Bioinformatics*, 01(October):1–18.
- [95] Down, T. A., Piipari, M., and Hubbard, T. J. P. (2011). Dalliance: interactive genome viewing on the web. *Bioinformatics*, 27(6):889–890.
- [96] Dua, P., Yoo, J., Kim, S., and Lee, D.-k. (2011). Modified siRNA structure with a single nucleotide bulge overcomes conventional siRNA-mediated off-target silencing. *Molecular Therapy*, 19(9):1676–1687.
- [97] Duggal, N. K. and Emerman, M. (2012). Evolutionary conflicts between viruses and restriction factors shape immunity. *Nature Reviews Immunology*, 12(10):687–695.
- [98] Editorial (2007). Democratizing proteomics data. *Nature Biotechnology*, 25(3):262.
- [99] Editorial (2014). Method of the year 2013. *Nature Methods*, 11(1):1.
- [100] Edlin, B. R. (2011). Perspective: test and treat this silent killer. *Nature*, 474(7350):S18–S19.
- [101] Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, 405(6788):823–826.
- [102] Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498.
- [103] Elbashir, S. M., Lendeckel, W., and Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development*, 15(2):188–200.
- [104] Elsik, C. G., Worley, K. C., Zhang, L., Milshina, N. V., Jiang, H., Reese, J. T., Childs, K. L., Venkatraman, A., Dickens, C. M., Weinstock, G. M., and Gibbs, R. A. (2006). Community annotation: procedures, protocols, and supporting tools. *Genome Research*, 16(11):1329–1333.

-
- [105] Emig, D., Blankenburg, H., Ramírez, F., and Albrecht, M. (2012). Functional characterization of human genes from exon expression and RNA interference results. In Larson, R. S., editor, *Bioinformatics and Drug Discovery*, volume 910 of *Methods in Molecular Biology*, pages 33–53. Humana Press, New York, NY, 2. ed. edition.
- [106] Emig, D., Sander, O., Mayr, G., and Albrecht, M. (2011). Structure collisions between interacting proteins. *PLoS ONE*, 6(6):e19581.
- [107] Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90.
- [108] Erfle, H., Neumann, B., Liebel, U., Rogers, P., Held, M., Walter, T., Ellenberg, J., and Pepperkok, R. (2007). Reverse transfection on cell arrays for high content screening microscopy. *Nature Protocols*, 2(2):392–399.
- [109] Espadaler, J., Romero-Isart, O., Jackson, R. M., and Oliva, B. (2005). Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, 21(16):3360–8.
- [110] Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O’Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figgeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, 3(89):89.
- [111] Eyckerman, S., Lemmens, I., Catteeuw, D., Verhee, A., Vandekerckhove, J., Lievens, S., and Tavernier, J. (2005). Reverse MAPPIT: screening for protein-protein interaction modifiers in mammalian cells. *Nature Methods*, 2(6):427–433.
- [112] Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616.
- [113] Fernandez-Garcia, M.-D., Mazzon, M., Jacobs, M., and Amara, A. (2009). Pathogenesis of flavivirus infections: using and abusing the host cell. *Cell Host & Microbe*, 5(4):318–328.
- [114] Fields, S. (2005). High-throughput two-hybrid analysis. The promise and the peril. *FEBS Journal*, 272(21):5391–5399.
- [115] Fields, S. and Song, O.-k. (1989). A novel genetic system to detect protein-protein interactions. *Nature*.

- [116] Finn, R. D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412.
- [117] Finn, R. D., Stalker, J. W., Jackson, D. K., Kulesha, E., Clements, J., and Pet-tett, R. (2007). ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, 23(12):1568–1570.
- [118] Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue):D281–D288.
- [119] Fire, A., Xu, S. S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–811.
- [120] Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.-A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.-P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rossé, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. (2005). Protein interaction mapping: a *Drosophila* case study. *Genome Research*, 15(3):376–384.
- [121] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue):D808–D815.
- [122] Friedman, A. and Perrimon, N. (2006). A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. *Nature*, 444(7116):230–234.
- [123] Frishman, D., Albrecht, M., Blankenburg, H., Bork, P., Harrington, E. D., Hermjakob, H., Jensen, L. J., Juan, D. A., Lengauer, T., Pagel, P., Schachter, V., and Valencia, A. (2009). Protein-protein interactions: analysis and prediction. In Frishman, D. and Valencia, A., editors, *Modern Genome Annotation - The Biosapiens Network*, chapter 6, pages 353–410. Springer, Wien, Austria.
- [124] Fukuhara, T. and Matsuura, Y. (2012). Role of miR-122 and lipid metabolism in HCV infection. *Journal of Gastroenterology*, 48(2):169–176.
- [125] Futschik, M. E., Chaurasia, G., and Herzog, H. (2007). Comparison of human protein-protein interaction maps. *Bioinformatics*, 23(5):605–611.

- [126] Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293.
- [127] Garai, Á., Zeke, A., Gögl, G., Tőro, I., Fördos, F., Blankenburg, H., Bárkai, T., Varga, J., Alexa, A., Emig, D., Albrecht, M., and Reményi, A. (2012). Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove. *Science Signaling*, 5(245):ra74.
- [128] Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Research*, 40(Web Server issue):W147–W151.
- [129] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636.
- [130] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.
- [131] Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E. L., Qiu, P., Bertelsen, A. H., Muir, A. J., Sulkowski, M., McHutchison, J. G., and Goldstein, D. B. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*, 461(7262):399–401.
- [132] Gel Moreno, B., Jenkinson, A. M., Jimenez, R. C., Messeguer Peypoch, X., and Hermjakob, H. (2011). easyDAS: automatic creation of DAS servers. *BMC Bioinformatics*, 12(1):23.
- [133] Gentleman, R. and Huber, W. (2007). Making the most of high-throughput protein-interaction data. *Genome Biology*, 8(10):112.

- [134] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.
- [135] Ghany, M. G., Strader, D. B., Thomas, D. L., and Seeff, L. B. (2009). Diagnosis, management, and treatment of hepatitis C: an update. *Hepatology*, 49(4):1335–1374.
- [136] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–36.
- [137] Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457.
- [138] Goel, R., Muthusamy, B., Pandey, A., and Prasad, T. S. K. (2011). Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Molecular Biotechnology*, 48(1):87–95.
- [139] Goldberg, D. S. and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8):4372–4376.
- [140] Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*, 24(15):1743–1744.
- [141] Grebely, J., Matthews, G. V., and Dore, G. J. (2011). Treatment of acute HCV infection. *Nature Reviews Gastroenterology & Hepatology*, 8(5):265–274.
- [142] Grigoriev, A. (2001). 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.

-
- [143] Guimarães, K. S., Jothi, R., Zotenko, E., and Przytycka, T. M. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biology*, 7(11):R104.
- [144] Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, 404(6775):293–296.
- [145] Hanash, S. (2002). The Human Proteome Organization: a mission to advance proteome knowledge. *Molecular & Cellular Proteomics*, 1(6):413–414.
- [146] Hao, L., He, Q., Wang, Z., Craven, M., Newton, M. A., and Ahlquist, P. (2013). Limited agreement of independent RNAi screens for virus-required host genes owes more to false-negative than false-positive factors. *PLoS Computational Biology*, 9(9):e1003235.
- [147] Hara, H., Aizaki, H., Matsuda, M., Shinkai-Ouchi, F., Inoue, Y., Murakami, K., Shoji, I., Kawakami, H., Matsuura, Y., Lai, M. M. C., Miyamura, T., Wakita, T., and Suzuki, T. (2009). Involvement of creatine kinase B in hepatitis C virus genome replication through interaction with the viral NS4A protein. *Journal of Virology*, 83(10):5137–5147.
- [148] Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120.
- [149] Helt, G. A., Nicol, J. W., Erwin, E., Blossom, E., Blanchard, S. G., Chervitz, S. A., Harmon, C., and Loraine, A. E. (2009). Genoviz Software Development Kit: Java tool kit for building genomics visualization applications. *BMC Bioinformatics*, 10:266.
- [150] Helt, G. A., Nicol, J. W., Erwin, E., Blossom, E., Blanchard, S. G., Chervitz, S. A., Harmon, C., and Loraine, A. E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *BMC Bioinformatics*, 10(20):266.
- [151] Herker, E., Harris, C., Hernandez, C., Carpentier, A., Kaehlcke, K., Rosenberg, A. R., Farese, R. V., and Ott, M. (2010). Efficient hepatitis C virus particle formation requires diacylglycerol acyltransferase-1. *Nature Medicine*, 16(11):1295–1298.
- [152] Hermjakob, H. (2006). The HUPO proteomics standards initiative—overcoming the fragmentation of proteomics data. *Proteomics*, 6 Suppl 2:34–38.
- [153] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B.,

- Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183.
- [154] Hirsch, A. J. (2010). The use of RNAi-based screens to identify host proteins involved in viral replication. *Future Microbiology*, 5(2):303–311.
- [155] Hoffmann, R. and Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2:ii252–ii258.
- [156] Hoffmann, T. W., Duverlie, G., Gilles, D., Bengrine, A., and Abderrahmane, B. (2012). MicroRNAs and hepatitis C virus: toward the end of miR-122 supremacy. *Virology Journal*, 9(1):109.
- [157] Hofmann, W. P. and Zeuzem, S. (2011). A new standard of care for the treatment of chronic HCV infection. *Nature Reviews Gastroenterology & Hepatology*, 8(5):257–264.
- [158] Hoofnagle, J. H. (2002). Course and outcome of hepatitis C. *Hepatology*, 36(5 Suppl 1):S21–S29.
- [159] Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730.
- [160] Horn, T., Sandmann, T., and Boutros, M. (2010). Design and evaluation of genome-wide libraries for RNAi screens. *Genome Biology*, 11(6):R61.
- [161] Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature Methods*, 8(4):341–346.
- [162] Hsieh, T.-Y., Matsumoto, M., Chou, H.-C., Schneider, R., Hwang, S. B., Lee, A. S., and Lai, M. M. C. (1998). Hepatitis C virus core protein interacts with heterogeneous nuclear ribonucleoprotein K. *Journal of Biological Chemistry*, 273(28):17651–17659.
- [163] Huang, H. and Bader, J. S. (2009). Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–378.
- [164] Huang, H., Barker, W. C., Chen, Y., and Wu, C. H. (2003). iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Research*, 31(1):390–392.

- [165] Huang, H., Jedynak, B. M., and Bader, J. S. (2007). Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3(11):e214.
- [166] Huang, T.-W., Tien, A.-C., Huang, W.-S., Lee, Y.-C. G., Peng, C.-L., Tseng, H.-H., Kao, C.-Y., and Huang, C.-Y. F. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–3276.
- [167] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. A., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pockock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41.
- [168] Huxley, H. (1965). The mechanism of muscular contraction. *Science*, 164(8):1356–1366.
- [169] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45.
- [170] Isserlin, R., El-Badrawi, R. A., and Bader, G. D. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database*, 2011:baq037.
- [171] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574.
- [172] Izarzugaza, J. M. G., Juan, D., Pons, C., Pazos, F., and Valencia, A. (2008). Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, 9:35.
- [173] Jackson, A., Bartz, S., Schelter, J., Kobayashi, S., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6):635–637.
- [174] Jackson, A. L., Burchard, J., Schelter, J., Chau, B. N., Cleary, M., Lim, L., and Linsley, P. S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA*, 12(7):1179–1187.
- [175] Jackson, A. L. and Linsley, P. S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9(1):57–67.

- [176] Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562.
- [177] Jang, H., Lim, J., Lim, J.-H., Park, S.-J., Lee, K.-C., and Park, S.-H. (2006). Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics*, 22(14):e220–e226.
- [178] Jansen, R. and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology*, 7(5):535–545.
- [179] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.
- [180] Jayapandian, M., Chapman, A., Tarcea, V. G., Yu, C., Elkiss, A., Ianni, A., Liu, B., Nandi, A., Santos, C., Andrews, P., Athey, B., States, D., and Jagadish, H. V. (2007). Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Research*, 35(Database issue):D566–D571.
- [181] Jenkinson, A. M., Albrecht, M., Birney, E., Blankenburg, H., Down, T. A., Finn, R. D., Hermjakob, H., Hubbard, T. J. P., Jimenez, R. C., Jones, P., Kähäri, A., Kulesha, E., Macías, J. R., Reeves, G. A., and Prlić, A. (2008). Integrating biological data—the Distributed Annotation System. *BMC Bioinformatics*, 9 Suppl 8:S3.
- [182] Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.
- [183] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.
- [184] Jones, D. M., Domingues, P., Targett-Adams, P., and McLauchlan, J. (2010). Comparison of U2OS and Huh-7 cells for identifying host factors that affect hepatitis C virus RNA replication. *The Journal of General Virology*, 91(Pt 9):2238–2248.
- [185] Jones, P., Vinod, N., Down, T. A., Hackmann, A., Kahari, A., Kretschmann, E., Quinn, A., Wieser, D., Hermjakob, H., and Apweiler, R. (2005). Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, 21(14):3198–3199.

- [186] Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of Molecular Biology*, 362(4):861–875.
- [187] Kamburov, A., Stelzl, U., and Herwig, R. (2012). IntScore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Research*, 40(Web Server issue):W140–W146.
- [188] Kato, T., Furusaka, A., Miyamoto, M., Date, T., Yasui, K., Hiramoto, J., Nagayama, K., Tanaka, T., and Wakita, T. (2001). Sequence analysis of hepatitis C virus isolated from a fulminant hepatitis patient. *Journal of Medical Virology*, 64(3):334–339.
- [189] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckof, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666.
- [190] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeifferberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2011). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database issue):D841–D846.
- [191] Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J. J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stümpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M. E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007). Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5:44.
- [192] Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A., Kinsella, R. J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A. J., and Yates, A. (2010). Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Research*, 38(Database issue):D563–D569.
- [193] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human

- Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–D772.
- [194] Keskin, O., Nussinov, R., and Gursoy, A. (2008). PRISM: protein-protein interaction prediction by structural matching. *Functional Proteomics*, 484:505–521.
- [195] Khattab, M.-A. (2009). Targeting host factors: A novel rationale for the management of hepatitis C virus. *World J. Gastroenterol.*, 15(28):3472.
- [196] Kiemer, L. and Cesareni, G. (2007). Comparative interactomics: comparing apples and pears? *Trends in Biotechnology*, 25(10):448–454.
- [197] Kim, S., Shin, S.-Y., Lee, I.-H., Kim, S.-J., Sriram, R., and Zhang, B.-T. (2008). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research*, 36(Web Server issue):W411–W415.
- [198] Kim, Y., Min, B., and Yi, G.-S. (2012). IDDI: integrated domain-domain interaction and protein interaction analysis system. *Proteome Science*, 10(Suppl 1):S9.
- [199] Klingström, T. and Plewczynski, D. (2010). Protein-protein interaction and pathway databases, a graphical review. *Briefings in Bioinformatics*, 12(6):702–713.
- [200] Koh, G. C. K. W., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. E. (2012). Analyzing protein-protein interaction networks. *Journal of Proteome Research*, 11(4):2014–2031.
- [201] König, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2011). Protein-RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(2):77–83.
- [202] König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Ireland, J. T., Chiang, C.-Y., Tu, B. P., Jesus, P. D. D., Lilley, C. E., Seidel, S., Opaluch, A. M., Caldwell, J. S., Weitzman, M. D., Kuhlen, K. L., Bandyopadhyay, S., Ideker, T., Orth, A. P., Miraglia, L. J., Bushman, F. D., Young, J. a., Chanda, S. K., Koenig, R., and De Jesus, P. D. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60.
- [203] Kritikos, G. D., Moschopoulos, C., Vazirgiannis, M., and Kossida, S. (2011). Noise reduction in protein-protein interaction graphs by the implementation of a novel weighting scheme. *BMC Bioinformatics*, 12(1):239.
- [204] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales,

- M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643.
- [205] Kronenberger, B. and Zeuzem, S. (2012). New developments in HCV therapy. *Journal of Viral Hepatitis*, 19 Suppl 1:48–51.
- [206] Kubrycht, J., Sigler, K., and Souček, P. (2012). Virtual interactomics of proteins from biochemical standpoint. *Molecular Biology International*, 2012:976385.
- [207] Kwong, A. D., Kauffman, R. S., Hurter, P., and Mueller, P. (2011). Discovery and development of telaprevir: an NS3-4A protease inhibitor for treating genotype 1 chronic hepatitis C virus. *Nature Biotechnology*, 29(11):993–1003.
- [208] Lalonde, S., Ehrhardt, D. W., Loqué, D., Chen, J., Rhee, S. Y., and Frommer, W. B. (2008). Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *The Plant Journal*, 53(4):610–635.
- [209] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati,

- R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- [210] Lanford, R. E., Hildebrandt-Eriksen, E. S., Petri, A., Persson, R., Lindow, M., Munk, M. E., Kauppinen, S., and Ørum, H. (2010). Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, 327(5962):198–201.
- [211] Launay, G. and Simonson, T. (2008). Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics*, 9:427.
- [212] Lederberg, B. J. and McCray, A. T. (2001). 'Ome sweet' omics— a genealogical treasury of words. *Scientist*, 15(7):8.
- [213] Lee, H., Deng, M., Sun, F., and Chen, T. (2006). An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 7:269.
- [214] Lee, I., Li, Z., and Marcotte, E. M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE*, 2(10):e988.

-
- [215] Lehner, B. and Fraser, A. G. (2004). A first-draft human protein-interaction map. *Genome Biology*, 5(9):R63.
- [216] Lengauer, T. (2012). Bioinformatical assistance of selecting anti-HIV therapies: where do we stand? *Intervirology*, 55(2):108–112.
- [217] Li, D., Liu, W., Liu, Z., Wang, J., Liu, Q., and Zhu, Y. (2008). PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Molecular & Cellular Proteomics*, 7(6):1043–1052.
- [218] Li, Q., Brass, A. L., Ng, A., Hu, Z., Xavier, R. J., Liang, T. J., and Elledge, S. J. (2009). A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38):16410–16415.
- [219] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543.
- [220] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardoza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(1):D857–D861.
- [221] Lievens, S., Eyckerman, S., Lemmens, I., and Tavernier, J. (2010). Large-scale protein interactome mapping: strategies and opportunities. *Expert Review of Proteomics*, 7(5):679–90.
- [222] Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabási, A.-L., Vidal, M., and Zoghbi, H. Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–814.
- [223] Lima, W. F., Prakash, T. P., Murray, H. M., Kinberger, G. A., Li, W., Chappell, A. E., Li, C. S., Murray, S. F., Gaus, H., Seth, P. P., Swayze, E. E., and Crooke, S. T. (2012). Single-stranded siRNAs activate RNAi in animals. *Cell*, 150(5):883–894.

- [224] Liu, Y., Liu, N., and Zhao, H. (2005). Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21(15):3279–3285.
- [225] Liu, Z.-P. and Chen, L. (2012). Proteome-wide prediction of protein-protein interactions from high-throughput data. *Protein & Cell*, 3(7):508–520.
- [226] Lopes, T. J. S., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J.-F., Neumann, G., Andrade-Navarro, M. A., Kawaoka, Y., and Kitano, H. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27(17):2414–2421.
- [227] Luo, Q., Pagel, P., Vilne, B., and Frishman, D. (2010). DIMA 3.0: Domain Interaction Map. *Nucleic Acids Research*, 39(Database issue):D724–D729.
- [228] Lupberger, J., Zeisel, M. B., Xiao, F., Thumann, C., Fofana, I., Zona, L., Davis, C., Mee, C. J., Turek, M., Gorke, S., Royer, C., Fischer, B., Zahid, M. N., Lavillette, D., Fresquet, J., Cosset, F.-L., Rothenberg, S. M., Pietschmann, T., Patel, A. H., Pessaux, P., Doffoël, M., Raffelsberger, W., Poch, O., McKeating, J. a., Brino, L., and Baumert, T. F. (2011). EGFR and EphA2 are host factors for hepatitis C virus entry and possible targets for antiviral therapy. *Nature Medicine*, 17(5):589–595.
- [229] Macías, J. R., Jiménez-Lozano, N., and Carazo, J. M. (2007). Integrating electron microscopy information into existing Distributed Annotation Systems. *Journal of Structural Biology*, 158(2):205–213.
- [230] Mackay, J. P., Sunde, M., Lowry, J. A., Crossley, M., and Matthews, J. M. (2007). Protein interactions: is seeing believing? *Trends in Biochemical Sciences*, 32(12):530–531.
- [231] Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175.
- [232] Mani, R., St Onge, R. P., Hartman, J. L., Giaever, G., and Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3461–3466.
- [233] Manns, M. P., Foster, G. R., Rockstroh, J. K., Zeuzem, S., Zoulim, F., and Houghton, M. (2007). The way forward in HCV treatment—finding the right path. *Nature Reviews Drug Discovery*, 6(12):991–1000.
- [234] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86.

- [235] Marine, S., Bahl, A., Ferrer, M., and Buehler, E. (2011). Common seed analysis to identify off-target effects in siRNA screens. *Journal of Biomolecular Screening*, 17(3):370–378.
- [236] Mathivanan, S., Periaswamy, B., Gandhi, T. K. B., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y. L., and Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5:S19.
- [237] Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126.
- [238] Matula, P., Kumar, A., Wörz, I., Erfle, H., Bartenschlager, R., Eils, R., and Rohr, K. (2009). Single-cell-based image analysis of high-throughput cell array screens for quantification of viral infection. *Cytometry Part A*, 75(4):309–318.
- [239] McDermott, J., Bumgarner, R., and Samudrala, R. (2005). Functional annotation from predicted protein interaction networks. *Bioinformatics*, 21(15):3217–26.
- [240] McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature*, 490(7422):138–142.
- [241] McMichael, A. J. and Haynes, B. F. (2012). Lessons learned from HIV-1 vaccine trials: new priorities and directions. *Nature Immunology*, 13(5):423–427.
- [242] Medina, I., Salavert, F., Sanchez, R., de Maria, A., Alonso, R., Escobar, P., Bleda, M., and Dopazo, J. (2013). Genome Maps, a new generation genome browser. *Nucleic Acids Research*, 41(Web Server issue):W41–W46.
- [243] Mering, C. V. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261.
- [244] Meyer, M. J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics*, 29(12):1577–1579.
- [245] Miyanari, Y., Atsuzawa, K., Usuda, N., Watashi, K., Hishiki, T., Zayas, M., Bartenschlager, R., Wakita, T., Hijikata, M., and Shimotohno, K. (2007). The lipid droplet is an important organelle for hepatitis C virus production. *Nature Cell Biology*, 9(9):1089–1097.
- [246] Moffat, J. and Sabatini, D. M. (2006). Building mammalian signalling pathways with RNAi screens. *Nature Reviews Molecular Cell Biology*, 7(3):177–187.

- [247] Mora, A. and Donaldson, I. M. (2011). iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics*, 12(1):455.
- [248] Moradpour, D., Penin, F., and Rice, C. M. (2007). Replication of hepatitis C virus. *Nature Reviews Microbiology*, 5(6):453–463.
- [249] Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536.
- [250] Mosca, R., Céol, A., and Aloy, P. (2012). Interactome3D: adding structural details to protein networks. *Nature Methods*, 10(1):47–53.
- [251] Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2013). 3Did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(Database issue):D374–D379.
- [252] Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? *Genome Research*, 11(12):1971–1973.
- [253] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C. P., Servant, F., and Sigrist, C. J. a. (2002). InterPro: an integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics*, 3(3):225–235.
- [254] Nagy, P. D. and Pogany, J. (2011). The dependence of viral RNA replication on co-opted host factors. *Nature Reviews Microbiology*, 10(2):137–149.
- [255] Naito, Y., Yoshimura, J., Morishita, S., and Ui-Tei, K. (2009). siDirect 2.0: updated software for designing functional siRNA with reduced seed-dependent off-target effect. *BMC Bioinformatics*, 10:392.
- [256] Navratil, V., de Chasse, B., Combe, C. R., and Lotteau, V. (2011). When the human viral infectome and disease networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Systems Biology*, 5(1):13.
- [257] Navratil, V., de Chasse, B., Meyniel, L., Delmotte, S., Gautier, C., André, P., Lotteau, V., and Roubourdin-Combe, C. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Research*, 37(Database issue):D661–D668.
- [258] Neduva, V. and Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Letters*, 579(15):3342–3345.

-
- [259] Negre, V. and Grunau, C. (2004). eL-DASionator: an LDAS upload file generator. *BMC Bioinformatics*, 5:55.
- [260] Ng, K.-L., Huang, C.-H., and Liu, H.-C. H.-C. (2008). Applications of domain-domain interactions in pathway study. *Computational Biology and Chemistry*, 32(2):81–87.
- [261] Ng, S.-K., Zhang, Z., and Tan, S.-H. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929.
- [262] Ng, T. I., Mo, H., Pilot-Matias, T., He, Y., Koev, G., Krishnan, P., Mondal, R., Pithawalla, R., He, W., Dekhtyar, T., Packer, J., Schurdak, M., and Molla, A. (2007). Identification of host genes involved in hepatitis C virus replication by small interfering RNA technology. *Hepatology*, 45(6):1413–1421.
- [263] Nguyen Ba, A. N., Yeh, B. J., van Dyk, D., Davidson, A. R., Andrews, B. J., Weiss, E. L., and Moses, A. M. (2012). Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science Signaling*, 5(215):rs1.
- [264] Nibbe, R. K., Chowdhury, S. A., Koyutürk, M., Ewing, R., and Chance, M. R. (2010). Protein-protein interaction networks and subnetworks in the biology of disease. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(3):357–367.
- [265] Nolte, A., Ott, K., Rohayem, J., Walker, T., Schlensak, C., and Wendel, H. P. (2012). Modification of small interfering RNAs to prevent off-target effects by the sense strand. *New Biotechnology*, 30(2):159–165.
- [266] Nooren, I. M. and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology*, 325(5):991–1018.
- [267] Olason, P. I. (2005). Integrating protein annotation resources through the Distributed Annotation System. *Nucleic Acids Research*, 33(Web Server issue):W468–W470.
- [268] Orchard, S., Albar, J.-P., Deutsch, E. W., Eisenacher, M., Binz, P.-A., and Hermjakob, H. (2010). Implementing data standards: a report on the HUPO-PSI workshop September 2009, Toronto, Canada. *Proteomics*, 10(10):1895–1898.
- [269] Orchard, S., Albar, J.-P., Deutsch, E. W., Eisenacher, M., Binz, P.-A., Martinez-Bartolome, S., Vizcaino, J. A., and Hermjakob, H. (2012). From proteomics data representation to public data flow: a report on the HUPO-PSI workshop September 2011, Geneva, Switzerland. *Proteomics*, 12(3):351–355.

- [270] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(Database issue):D358–D363.
- [271] Orchard, S., Binz, P.-A., Jones, A. R., Vizcaino, J. A., Deutsch, E. W., and Hermjakob, H. (2013). Preparing to work with big data in proteomics - a report on the HUPO-PSI Spring Workshop: April 15-17, 2013, Liverpool, UK. *Proteomics*, 13(20):2931–2937.
- [272] Orchard, S., Deutsch, E. W., Binz, P.-A., Jones, A. R., Creasy, D., Montechi-Palazzi, L., Corthals, G., and Hermjakob, H. (2009). Annual spring meeting of the Proteomics Standards Initiative. *Proteomics*, 9(19):4429–4432.
- [273] Orchard, S. and Hermjakob, H. (2011). Data standardization by the HUPO-PSI: how has the community benefitted? In Hamacher, M., Eisenacher, M., and Stephan, C., editors, *Data Mining in Proteomics*, volume 696 of *Methods in Molecular Biology*, pages 149–160. Humana Press.
- [274] Orchard, S. and Hermjakob, H. (2011). Preparing molecular interaction data for publication. In Wu, C. H. and Chen, C., editors, *Bioinformatics for Comparative Proteomics*, volume 694 of *Methods in Molecular Biology*, pages 229–236. Humana Press.
- [275] Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics*, 3(7):1374–1376.
- [276] Orchard, S., Hermjakob, H., Taylor, C., Binz, P.-A., Hoogland, C., Julian, R., Garavelli, J. S., Aebersold, R., and Apweiler, R. (2006). Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4-6, 2005. *Proteomics*, 6(3):738–741.
- [277] Orchard, S. and Kerrien, S. (2010). Molecular interactions and data standardisation. In Hubbard, S. J. and Jones, A. R., editors, *Proteome Bioinformatics*, volume 604 of *Methods in Molecular Biology*, pages 309–318. Humana Press, Totowa, NJ.
- [278] Orchard, S., Kerrien, S., Abbani, S., and Aranda, B. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(4):345–350.

- [279] Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothrin, J., and Hermjakob, H. (2007). Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, 7 Suppl 1:28–34.
- [280] Orchard, S., Montecchi-Palazzi, L., Deutsch, E. W., Binz, P.-A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., and Hermjakob, H. (2007). Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23-25, 2007 Ecole Nationale Supérieure (ENS), Lyon, France. *Proteomics*, 7(19):3436–3440.
- [281] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25(8):894–898.
- [282] Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stümpflen, V., and Frishman, D. (2008). DIMA 2.0—predicted and known domain interactions. *Nucleic Acids Research*, 36(Database issue):D651–D655.
- [283] Parfi, A., Jaroszewicz, J., and Flisiak, R. (2007). Specifically targeted antiviral therapy for hepatitis C virus. *Journal of Gastroenterology*, 13(43):5673–5681.
- [284] Patil, A., Nakai, K., and Nakamura, H. (2011). HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Research*, 39(Database issue):D744–D749.
- [285] Patil, A. and Nakamura, H. (2005). Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6:100.
- [286] Pawson, T. and Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. *Genes & Development*, 14(9):1027–1047.
- [287] Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452.
- [288] Pelkmans, L., Fava, E., Grabner, H., Hannus, M., Habermann, B., Krausz, E., and Zerial, M. (2005). Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature*, 436(7047):78–86.

- [289] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285–4288.
- [290] Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 Suppl 4:S21.
- [291] Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443.
- [292] Petzko, G. A. and Ringe, D. (2004). *Protein Structure and Function*. Primers in biology. New Science Press.
- [293] Pichlmair, A., Kandasamy, K., and Alvisi, G. (2012). Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature*, 487(7408):486–490.
- [294] Plewczynski, D. and Ginalski, K. (2009). The interactome: predicting the protein-protein interactions in cells. *Cellular & Molecular Biology Letters*, 14(1):1–22.
- [295] Plewczynski, D. and Klingström, T. (2011). GIDMP: good protein-protein interaction data metamining practice. *Cellular & Molecular Biology Letters*, 16(2):258–263.
- [296] Poenisch, M. and Bartenschlager, R. (2010). New insights into structure and replication of the hepatitis C virus and clinical implications. *Seminars in Liver Disease*, 30(4):333–347.
- [297] Poordad, F. and Dieterich, D. (2012). Treating hepatitis C: current standard of care and emerging direct-acting antiviral agents. *Journal of Viral Hepatitis*, 19(7):449–464.
- [298] Prieto, C. and De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Research*, 34(Web Server issue):W298–W302.
- [299] Prlić, A., Birney, E., Cox, T., Down, T. A., Finn, R., Gräf, S., Jackson, D., Kähäri, A., Kulesha, E., Pettett, R., Smith, J., Stalker, J., and Hubbard, T. J. P. (2006). The distributed annotation system for integration of biological data. In *Proceedings of the Third International Conference on Data Integration in the Life Sciences*, DILS’06, pages 195–203, Berlin, Heidelberg. Springer-Verlag.

-
- [300] Prlić, A., Down, T. A., and Hubbard, T. J. P. (2005). Adding some SPICE to DAS. *Bioinformatics*, 21 Suppl 2:ii40–ii41.
- [301] Prlić, A., Down, T. A., Kulesha, E., Finn, R. D., Kähäri, A., and Hubbard, T. J. P. (2007). Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, 8:333.
- [302] Prlić, A., Yates, A., Bliven, S., and Rose, P. (2012). BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, 28(20):2693–2695.
- [303] Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.
- [304] Przulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348.
- [305] Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it’s about time. *Briefings in Bioinformatics*, 11(1):15–29.
- [306] Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, 39(11):1338–1349.
- [307] Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R. (2008). DOMINE: a database of protein domain interactions. *Nucleic Acids Research*, 36(Database issue):D656–D661.
- [308] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409(6817):211–215.
- [309] Rajagopala, S. V., Goll, J., Gowda, N. D. D., Sunil, K. C., Titz, B., Mukherjee, A., Mary, S. S., Raviswaran, N., Poojari, C. S., Ramachandra, S., Shtivelband, S., Blazie, S. M., Hofmann, J., and Uetz, P. (2008). MPI-LIT: a literature-curated dataset of microbial binary protein–protein interactions. *Bioinformatics*, 24(22):2622–2627.
- [310] Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):R40.

- [311] Ramirez, F. (2011). *Novel Approaches to the Integration and Analysis of Systems Biology Data*. PhD thesis, Universität des Saarlandes.
- [312] Ramírez, F., Lawyer, G., and Albrecht, M. (2011). Novel search method for the discovery of functional relationships. *Bioinformatics*, 28(2):269–276.
- [313] Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., and Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552.
- [314] Randall, G., Panis, M., Cooper, J. D., Tellinghuisen, T. L., Sukhodolets, K. E., Pfeffer, S., Landthaler, M., Landgraf, P., Kan, S., Lindenbach, B. D., Chien, M., Weir, D. B., Russo, J. J., Ju, J., Brownstein, M. J., Sheridan, R., Sander, C., Zavolan, M., Tuschl, T., and Rice, C. M. (2007). Cellular cofactors affecting hepatitis C virus infection and replication. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12884–12889.
- [315] Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.
- [316] Rebholz-Schuhmann, D., Oellrich, A., and Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, 13(12):829–839.
- [317] Reeves, G. A., Eilbeck, K., Magrane, M., Donovan, C. O., Montec, L., Harris, M. A., Orchard, S., Jimenez, R. C., Prlic, A., Tim, J. P., Hermjakob, H., and Thornton, J. M. (2008). The Protein Feature Ontology: a tool for the unification of protein feature annotations. *Bioinformatics*, 24(23):2767–2772.
- [318] Regenmortel, M. H. V. V. (2004). Reductionism and complexity in molecular biology. *EMBO reports*, 5(11):1016–1020.
- [319] Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskaya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *Journal of Biology*, 5(4):11.
- [320] Reimand, J., Hui, S., Jain, S., Law, B., and Bader, G. D. (2012). Domain-mediated protein interaction prediction: From genome to network. *FEBS Letters*, 586(17):2751–2763.
- [321] Reiss, S., Rebhan, I., Backes, P., Romero-Brey, I., Erfle, H., Matula, P., Kaderali, L., Poenisch, M., Blankenburg, H., Hiet, M.-S., Longerich, T., Diehl, S.,

- Ramirez, F., Balla, T., Rohr, K., Kaul, A., Bühler, S., Pepperkok, R., Lengauer, T., Albrecht, M., Eils, R., Schirmacher, P., Lohmann, V., and Bartenschlager, R. (2011). Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell Host & Microbe*, 9(1):32–45.
- [322] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23(8):951–955.
- [323] Rieber, N., Knapp, B., Eils, R., and Kaderali, L. (2009). RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens. *Bioinformatics*, 25(5):678–679.
- [324] Rigaut, G., Shevchenko, A., Rutz, B., and Wilm, M. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(October):7–9.
- [325] Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, 6(10):R89.
- [326] Rosen, H. R. (2011). Clinical practice. Chronic hepatitis C infection. *The New England Journal of Medicine*, 364(25):2429–2438.
- [327] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamasas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.
- [328] Ruffner, H., Bauer, A., and Bouwmeester, T. (2007). Human protein-protein interaction networks and the value for drug discovery. *Drug Discovery Today*, 12(17-18):709–716.
- [329] Salazar, G., García, L., Jones, P., and Jimenez, R. (2012). MyDas, an extensible Java DAS server. *PloS ONE*, 7(9):e44180.
- [330] Salazar, G. A., Jimenez, R. C., Garcia, A., Hermjakob, H., Mulder, N., and Blake, E. (2011). DAS Writeback: A Collaborative Annotation System. *BMC Bioinformatics*, 12(1):143.

- [331] Salloum, S. and Tai, A. W. (2012). Treating hepatitis C infection by targeting the host. *Translational Research*, 159(6):421–429.
- [332] Salwinski, L., Licata, L., Winter, A., Thorneycroft, D., Khadake, J., Ceol, A., Aryamontri, A. C., Oughtred, R., Livstone, M., Boucher, L., Botstein, D., Dolinski, K., Berardini, T., Huala, E., Tyers, M., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2009). Recurated protein interaction datasets. *Nature Methods*, 6(12):860–861.
- [333] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–D451.
- [334] Samreen, B., Khaliq, S., Ashfaq, U. A., Khan, M., Afzal, N., Shahzad, M. A., Riaz, S., and Jahan, S. (2012). Hepatitis C virus entry: role of host and viral factors. *Infection, Genetics and Evolution*, 12(8):1699–1709.
- [335] Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., Rechenmann, F., and Jacq, B. (1999). Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Research*, 27(1):89–94.
- [336] Satagopam, V. P., Theodoropoulou, M. C., Stampolakis, C. K., Pavlopoulos, G. A., Papandreou, N. C., Bagos, P. G., Schneider, R., and Hamodrakas, S. J. (2010). GPCRs, G-proteins, effectors and their interactions: human-gpDB, a database employing visualization tools and data integration techniques. *Database*, 2010:baq019.
- [337] Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PloS ONE*, 7(2):e31826.
- [338] Schaefer, M. H., Lopes, T. J. S., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J.-F., Louis-Jeune, C., Einfeld, A. J., Neumann, G., Perez-Iratxeta, C., Kawaoka, Y., Kitano, H., and Andrade-Navarro, M. A. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Computational Biology*, 9(1):e1002860.
- [339] Schelhorn, S.-E., Mestre, J., Albrecht, M., and Zotenko, E. (2011). Inferring physical protein contacts from large-scale purification data of protein complexes. *Molecular & Cellular Proteomics*, 10(6):M110.004929.
- [340] Schlicker, A. and Albrecht, M. (2008). FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(Database issue):D434–D439.

-
- [341] Schlicker, A. and Albrecht, M. (2010). FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Research*, 38(Database issue):D244–D248.
- [342] Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302.
- [343] Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–865.
- [344] Schmidt, E. E., Pelz, O., Buhlmann, S., Kerr, G., Horn, T., and Boutros, M. (2013). GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Research*, 41(D1):D1021–6.
- [345] Scholtens, D., Vidal, M., and Gentleman, R. (2005). Local modeling of global interactome networks. *Bioinformatics*, 21(17):3548–3557.
- [346] Shah, S. P., Huang, Y., Xu, T., Yuen, M. M. S., Ling, J., and Ouellette, B. F. F. (2005). Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, 6:34.
- [347] Sharma, S. and Rao, A. (2009). RNAi screening: tips and techniques. *Nature Immunology*, 10(8):799–804.
- [348] Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology*, 3(3):e42.
- [349] Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3(4):e43.
- [350] Sigoillot, F. D., Lyman, S., Huckins, J. F., Adamson, B., Chung, E., Quatrocchi, B., and King, R. W. (2012). A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. *Nature Methods*, 9(4):363–366.
- [351] Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H., and Williams, B. R. G. (2003). Activation of the interferon system by short-interfering RNAs. *Nature Cell Biology*, 5(9):834–849.
- [352] Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

- [353] Snijder, B., Sacher, R., Rämö, P., Damm, E.-M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263):520–523.
- [354] Speicher, N. K. (2010). Network analysis of viral host factors. Bachelor thesis, Universität des Saarlandes, Saarbrücken.
- [355] Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923.
- [356] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue):D535–D539.
- [357] Stein, A. and Aloy, P. (2008). Contextual specificity in peptide-mediated protein interactions. *PloS ONE*, 3(7):e2524.
- [358] Stein, A., Céol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 39(Database issue):D718–D723.
- [359] Stein, L. (2002). Creating a bioinformatics nation. *Nature*, 417(6885):119–120.
- [360] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlauff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968.
- [361] Strader, D. B. and Seeff, L. B. (2012). A brief history of the treatment of viral hepatitis C. *Clinical Liver Disease*, 1(1):6–11.
- [362] Strömbäck, L. and Lambrix, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–4407.
- [363] Stumpf, M. P. H., Thorne, T., Silva, E. D., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6959–6964.
- [364] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067.

- [365] Supekova, L., Supek, F., Lee, J., Chen, S., Gray, N., Pezacki, J. P., Schlapbach, A., and Schultz, P. G. (2008). Identification of human kinases involved in hepatitis C virus replication by small interference RNA library screening. *The Journal of Biological Chemistry*, 283(1):29–36.
- [366] Suppiah, V., Moldovan, M., Ahlenstiel, G., Berg, T., Weltman, M., Abate, M. L., Bassendine, M., Spengler, U., Dore, G. J., Powell, E., Riordan, S., Sheridan, D., Smedile, A., Fragomeli, V., Müller, T., Bahlo, M., Stewart, G. J., Booth, D. R., and George, J. (2009). IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nature Genetics*, 41(10):1100–1104.
- [367] Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. (2006). A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360.
- [368] Syed, G. H., Amako, Y., and Siddiqui, A. (2010). Hepatitis C virus hijacks host lipid metabolism. *Trends in Endocrinology and Metabolism*, 21(1):33–40.
- [369] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., Doerks, T., Stark, M., Müller, J., Bork, P., Jensen, L. J., von Mering, C., and Mering, C. v. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):D561–D568.
- [370] Tai, A. W., Benita, Y., Peng, L. F., Kim, S.-S., Sakamoto, N., Xavier, R. J., and Chung, R. T. (2009). A functional genomic screen identifies cellular cofactors of hepatitis C virus replication. *Cell Host & Microbe*, 5(3):298–307.
- [371] Tanaka, Y., Nishida, N., Sugiyama, M., Kurosaki, M., Matsuura, K., Sakamoto, N., Nakagawa, M., Korenaga, M., Hino, K., Hige, S., Ito, Y., Mita, E., Tanaka, E., Mochida, S., Murawaki, Y., Honda, M., Sakai, A., Hiasa, Y., Nishiguchi, S., Koike, A., Sakaida, I., Imamura, M., Ito, K., Yano, K., Masaki, N., Sugauchi, F., Izumi, N., Tokunaga, K., and Mizokami, M. (2009). Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nature Genetics*, 41(10):1105–1109.
- [372] Tang, H. and Grisé, H. (2009). Cellular and molecular biology of HCV infection and hepatitis. *Clinical Science*, 117(2):49–65.
- [373] Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204.
- [374] Thakur, N., Qureshi, A., and Kumar, M. (2011). VIRsiRNadb: a curated database of experimentally validated viral siRNA/shRNA. *Nucleic Acids Research*, 40(Database issue):D230–D236.

- [375] The Uniprot Consortium (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41(Database issue):D43–D47.
- [376] Thomas, D. L., Thio, C. L., Martin, M. P., Qi, Y., Ge, D., O’Huigin, C., Kidd, J., Kidd, K., Khakoo, S. I., Alexander, G., Goedert, J. J., Kirk, G. D., Donfield, S. M., Rosen, H. R., Tobler, L. H., Busch, M. P., McHutchison, J. G., Goldstein, D. B., and Carrington, M. (2009). Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature*, 461(7265):798–801.
- [377] Thornton, J. (2009). Annotations for all by all - the BioSapiens network. *Genome Biology*, 10(2):401.
- [378] Trabuco, L. G., Betts, M. J., and Russell, R. B. (2012). Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, 58(4):343–348.
- [379] Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Olason, P. I., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. A., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S. r., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., Orengo, C. A., Patthy, L., Thornton, J. M., Tramontano, A., and Valencia, A. (2007). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13):5495–5500.
- [380] Trotard, M., Lepère-Douard, C., Régeard, M., Piquet-Pellorce, C., Lavillette, D., Cosset, F.-L., Gripon, P., and Le Seyec, J. (2009). Kinases required in hepatitis C virus entry and replication highlighted by small interference RNA screening. *The FASEB Journal*, 23(11):3780–3789.
- [381] Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database*, 2010:baq026.
- [382] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627.

- [383] Vaillancourt, F. H., Pilote, L., Cartier, M., Lippens, J., Liuzzi, M., Bethell, R. C., Cordingley, M. G., and Kukulj, G. (2009). Identification of a lipid kinase as a host factor involved in hepatitis C virus RNA replication. *Virology*, 387(1):5–10.
- [384] Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3):368–373.
- [385] van den Berg, B. H. J., McCarthy, F. M., Lamont, S. J., and Burgess, S. C. (2010). Re-annotation is an essential step in systems biology modeling of functional genomics data. *PLoS ONE*, 5(5):e10642.
- [386] van Haagen, H. H. H. B. M., 't Hoen, P. A. C., de Morrée, A., van Roon-Mom, W. M. C., Peters, D. J., Roos, M., Mons, B., van Ommen, G.-J., and Schuemie, M. J. (2011). In silico discovery and experimental validation of new protein-protein interactions. *Proteomics*, 11(5):843–853.
- [387] Veiga, D. F. T., Deus, H. F., Akdemir, C., Vasconcelos, A. T. R., and Almeida, J. S. (2009). DASMiner: discovering and integrating data from DAS sources. *BMC Systems Biology*, 17(3):109.
- [388] Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., Smet, A.-S. D., Dann, E., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Vidal, M., and Wanker, E. E. (2009). An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90.
- [389] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L.,

- Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., Mccawley, S., Mcintosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., Mcdaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(February):1304–1351.
- [390] Vermehren, J. and Sarrazin, C. (2011). New HCV therapies on the horizon. *Clinical Microbiology and Infection*, 17(2):122–134.
- [391] Vidal, M. (2005). Interactome modeling. *FEBS Letters*, 579(8):1834–1838.
- [392] Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–98.
- [393] Villaveces, J. M., Jimenez, R. C., Garcia, L. J., Salazar, G. A., Gel, B., Mulder, N., Martin, M., Garcia, A., and Hermjakob, H. (2011). Dasty3, a web framework for DAS. *Bioinformatics*, 27(18):2616–2617.
- [394] von Hahn, T. and Rice, C. M. (2008). Hepatitis C virus entry. *The Journal of Biological Chemistry*, 283(7):3689–3693.
- [395] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. a., and Bork, P. (2005). STRING: known and pre-

- dicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–D437.
- [396] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.
- [397] Škunca, N., Altenhoff, A., and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, 8(5):e1002533.
- [398] Wakita, T., Pietschmann, T., Kato, T., and Date, T. (2005). Production of infectious hepatitis C virus in tissue culture from a cloned viral genome. *Nature Medicine*, 11(7):791–796.
- [399] Walhout, A. J. and Vidal, M. (2001). Protein interaction maps for model organisms. *Nature Reviews Molecular Cell Biology*, 2(1):55–62.
- [400] Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287(5450):116–122.
- [401] Wang, R.-S., Wang, Y., Wu, L.-Y., Zhang, X.-S., and Chen, L. (2007). Analysis on multi-domain cooperation for predicting protein-protein interactions. *BMC Bioinformatics*, 8:391.
- [402] Wass, M. N., David, A., and Sternberg, M. J. E. (2011). Challenges for the prediction of macromolecular interactions. *Current Opinion in Structural Biology*, 21(3):382–390.
- [403] Wass, M. N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology*, 7(469):1–8.
- [404] Welsch, C. and Zeuzem, S. (2012). Clinical relevance of HCV antiviral drug resistance. *Current Opinion in Virology*, 2(5):651–655.
- [405] Welsch, S., Miller, S., Romero-Brey, I., Merz, A., Bleck, C. K. E., Walther, P., Fuller, S. D., Antony, C., Krijnse-Locker, J., and Bartenschlager, R. (2009). Composition and three-dimensional architecture of the dengue virus replication and assembly sites. *Cell Host & Microbe*, 5(4):365–375.
- [406] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D.,

- Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Database issue):D13–21.
- [407] Wodak, S., Vlasblom, J., Turinsky, A., and Pu, S. (2013). Protein-protein interaction networks: the puzzling riches. *Current Opinion in Structural Biology*, 23(6):941–953.
- [408] Woerz, I., Lohmann, V., and Bartenschlager, R. (2009). Hepatitis C virus replicons: dinosaurs still in business? *Journal of Viral Hepatitis*, 16(1):1–9.
- [409] Wuchty, S., Barabási, A.-L., and Ferdig, M. T. (2006). Stable evolutionary signal in a yeast protein interaction network. *BMC Evolutionary Biology*, 6:8.
- [410] Xenarios, I. and Eisenberg, D. (2001). Protein interaction databases. *Current Opinion in Biotechnology*, 12(4):334–349.
- [411] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291.
- [412] Xia, K., Dong, D., and Han, J.-D. J. (2006). IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, 7:508.
- [413] Xue, Q., Ding, H., Liu, M., Zhao, P., Gao, J., Ren, H., Liu, Y., and Qi, Z. T. (2007). Inhibition of hepatitis C virus replication and expression by small interfering RNA targeting host cellular genes. *Archives of Virology*, 152(5):955–962.
- [414] Yan, Y. and Marriott, G. (2003). Analysis of protein interactions using fluorescence technologies. *Current Opinion in Chemical Biology*, 7(5):635–640.
- [415] Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, 39(Database issue):D730–D735.
- [416] You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, 26(21):2744–2751.
- [417] Yu, D., Pendergraft, H., Liu, J., Kordasiewicz, H. B., Cleveland, D. W., Swayze, E. E., Lima, W. F., Crooke, S. T., Prakash, T. P., and Corey, D. R. (2012). Single-stranded RNAs use RNAi to potently and allele-selectively inhibit mutant huntingtin expression. *Cell*, 150(5):895–908.

- [418] Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.
- [419] Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, 14(6):1107–1118.
- [420] Yu, J. and Finley, R. L. (2009). Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics*, 25(1):105–111.
- [421] Yu, J., Pacifico, S., Liu, G., and Finley, R. L. (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9:461.
- [422] Yuan, Z., Wu, X., Liu, C., Xu, G., and Wu, Z. (2012). Asymmetric siRNA: new strategy to improve specificity and reduce off-target gene expression. *Human Gene Therapy*, 86(25):1–41.
- [423] Zanon, A., Rakovic, A., Blankenburg, H., Doncheva, N. T., Schwienbacher, C., Serafin, A., Alexa, A., Weichenberger, C. X., Albrecht, M., Klein, C., Hicks, A. A., Pramstaller, P. P., Domingues, F. S., and Pichler, I. (2013). Profiling of Parkin-Binding Partners Using Tandem Affinity Purification. *PloS ONE*, 8(11):e78648.
- [424] Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C., and Weinstein, J. N. (2004). Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 5(1):80.
- [425] Zeuzem, S., Andreone, P., Pol, S., Lawitz, E., Diago, M., Roberts, S., Focaccia, R., Younossi, Z., Foster, G. R., Horban, A., Ferenci, P., Nevens, F., Müllhaupt, B., Pockros, P., Terg, R., Shouval, D., van Hoek, B., Weiland, O., Van Heeswijk, R., De Meyer, S., Luo, D., Boogaerts, G., Polo, R., Picchio, G., and Beumont, M. (2011). Telaprevir for retreatment of HCV infection. *The New England Journal of Medicine*, 364(25):2417–2428.
- [426] Zhang, L. V., Wong, S. L., King, O. D., and Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5:38.

- [427] Zhang, Q., Petrey, D., Deng, L., Qiang, L., and Shi, Y. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560.
- [428] Zhang, X. D. (2007). A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*, 89(4):552–561.
- [429] Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., Swearingen, V., Yildirim, M. A., Yan, H., Dricot, A., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., Dupuy, D., Brasseur, R., Hill, D. E., Cusick, M. E., and Vidal, M. (2009). Edgetic perturbation models of human inherited disorders. *Molecular Systems Biology*, 5:321.
- [430] Zhou, H., Xu, M., Huang, Q., Gates, A. T., Zhang, X. D., Castle, J. C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D. J., and Espeseth, A. S. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host & Microbe*, 4(5):495–504.
- [431] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M., and Snyder, M. (2001). Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–2105.

Appendix

1 Technical documents

HUPO PSI MI 2.5 document formats

The following examples contain different representations of a binary protein-protein interaction between two *Arabidopsis thaliana* proteins. The representation as HUPO-PSI-MI XML 2.5 is shown in Listing 1. The same interaction as MITAB2.5 is shown in Listing 2 and as MITAB2.7 representation in Listing 3. The interaction data in the examples have been retrieved from the IntAct website (<http://www.ebi.ac.uk/intact/>) on 14 December 2014.

```

1 <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2 <entrySet minorVersion="4" version="5" level="2" xsi:schemaLocation="http://psi.hupo.org/mi/mif http://psidev.sourceforge.net/mi/
  rel25/src/MIF254.xsd" xmlns="http://psi.hupo.org/mi/mif" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
3   <entry>
4     <source releaseDate="2013-05-03+01:00">
5       <names>
6         <shortLabel>UniProt</shortLabel>
7         <fullName>UniProt</fullName>
8       </names>
9       <xref>
10        <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0486" dbAc="MI:0488" db="psi-mi"/>
11        <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-3989247" dbAc="MI:0469" db="intact"/>
12      </xref>
13    </source>
14    <experimentList>
15      <experimentDescription id="1441894">
16        <names>
17          <shortLabel>el_din-2004-1</shortLabel>
18          <fullName>DISTORTED2 encodes an ARPC2 subunit of the putative Arabidopsis ARP2/3 complex.</fullName>
19        </names>
20        <bibref>
21          <xref>
22            <primaryRef refTypeAc="MI:0358" refType="primary-reference" id="15086808" dbAc="MI:0446" db="pubmed"/>
23          </xref>
24        </bibref>
25        <xref>
26          <primaryRef refTypeAc="MI:0356" refType="identity" id="EBI-6620106" dbAc="MI:0469" db="intact"/>
27          <secondaryRef refTypeAc="MI:0662" refType="imex-primary" id="IM-18781" dbAc="MI:0670" db="imex"/>
28        </xref>
29      <hostOrganismList>
30        <hostOrganism ncbiTaxId="4932">
31          <names>
32            <shortLabel>yeasx</shortLabel>
33            <fullName>Saccharomyces cerevisiae (Baker's yeast)</fullName>
34          </names>
35        </hostOrganism>
36      </hostOrganismList>
37      <interactionDetectionMethod>
38        <names>
39          <shortLabel>2 hybrid</shortLabel>
40          <fullName>two hybrid</fullName>
41          <alias type="go synonym" typeAc="MI:0303">2H</alias>
42          <alias type="go synonym" typeAc="MI:0303">2h</alias>
43          <alias type="go synonym" typeAc="MI:0303">Gal4 transcription regeneration</alias>
44          <alias type="go synonym" typeAc="MI:0303">two-hybrid</alias>
45          <alias type="go synonym" typeAc="MI:0303">yeast two hybrid</alias>
46          <alias type="go synonym" typeAc="MI:0303">2-hybrid</alias>
47          <alias type="go synonym" typeAc="MI:0303">classical two hybrid</alias>
48          <alias type="synonym" typeAc="MI:1041">Y2H</alias>
49        </names>
50        <xref>
51          <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0018" dbAc="MI:0488" db="psi-mi"/>
52          <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="1946372" dbAc="MI:0446" db="pubmed"/>
53          <secondaryRef refTypeAc="MI:0357" refType="method reference" id="10967325" dbAc="MI:0446" db="pubmed"/>
54          <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-94" dbAc="MI:0469" db="intact"/>
55          <secondaryRef refTypeAc="MI:0357" refType="method reference" id="12634794" dbAc="MI:0446" db="pubmed"/>
56        </xref>
57      </interactionDetectionMethod>
58    <participantIdentificationMethod>

```

```

59     <names>
60       <shortLabel>predetermined</shortLabel>
61       <fullName>predetermined participant</fullName>
62     </names>
63     <xref>
64       <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0396" dbAc="MI:0488" db="psi-mi"/>
65       <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-1465" dbAc="MI:0469" db="intact"/>
66       <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
67     </xref>
68   </participantIdentificationMethod>
69   <attributeList>
70     <attribute nameAc="MI:0634" name="contact-email">dszyman@purdue.edu</attribute>
71     <attribute nameAc="MI:0636" name="author-list">El-Din El-Assal S., Le J., Basu D., Mallery E.L., Szymanski D.B.</attribute>
72     <attribute nameAc="MI:0885" name="journal">Plant J. (0960-7412)</attribute>
73     <attribute nameAc="MI:0886" name="publication year">2004</attribute>
74     <attribute nameAc="MI:0955" name="curation depth">imex curation</attribute>
75     <attribute nameAc="MI:0957" name="full coverage">Only protein-protein interactions</attribute>
76     <attribute nameAc="MI:0959" name="imex curation"/>
77   </attributeList>
78 </experimentDescription>
79 </experimentList>
80 <interactorList>
81   <interactor id="1441895">
82     <names>
83       <shortLabel>arc2a_arath</shortLabel>
84       <fullName>Actin-related protein 2/3 complex subunit 2A</fullName>
85       <alias type="gene name" typeAc="MI:0301">ARPC2A</alias>
86       <alias type="gene name synonym" typeAc="MI:0302">Actin-related protein C2A</alias>
87       <alias type="gene name synonym" typeAc="MI:0302">Arp2/3 complex 34 kDa subunit</alias>
88       <alias type="gene name synonym" typeAc="MI:0302">DIS2</alias>
89       <alias type="gene name synonym" typeAc="MI:0302">Protein DISTORTED TRICHOMES 2</alias>
90       <alias type="orf name" typeAc="MI:0306">T17H7.13</alias>
91       <alias type="locus name" typeAc="MI:0305">At1g30825</alias>
92     </names>
93     <xref>
94       <primaryRef refTypeAc="MI:0356" refType="identity" version="TrEMBL_19" id="Q8LGI3" dbAc="MI:0486" db="uniprotkb"/>
95       <secondaryRef refTypeAc="MI:0360" refType="secondary-ac" version="SP_68" id="Q9SY27" dbAc="MI:0486" db="uniprotkb"/>
96       <secondaryRef id="GO:0005737" dbAc="MI:0448" db="go"/>
97       <secondaryRef id="GO:0005885" dbAc="MI:0448" db="go"/>
98       <secondaryRef id="GO:0042995" dbAc="MI:0448" db="go"/>
99       <secondaryRef id="GO:0007015" dbAc="MI:0448" db="go"/>
100      <secondaryRef id="GO:0030833" dbAc="MI:0448" db="go"/>
101      <secondaryRef id="GO:0010090" dbAc="MI:0448" db="go"/>
102      <secondaryRef id="IPI00521701" dbAc="MI:0675" db="ipi"/>
103      <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-1547736" dbAc="MI:0469" db="intact"/>
104      <secondaryRef id="IPRO07188" dbAc="MI:0449" db="interpro"/>
105      <secondaryRef id="IPI00521322" dbAc="MI:0675" db="ipi"/>
106      <secondaryRef id="NP_564364.1" dbAc="MI:0481" db="refseq"/>
107    </xref>
108    <interactorType>
109      <names>
110        <shortLabel>protein</shortLabel>
111        <fullName>protein</fullName>
112      </names>
113      <xref>
114        <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0326" dbAc="MI:0488" db="psi-mi"/>
115        <secondaryRef refTypeAc="MI:0361" refType="see-also" id="SO:0000358" dbAc="MI:0601" db="so"/>
116        <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-619654" dbAc="MI:0469" db="intact"/>
117        <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
118      </xref>
119    </interactorType>
120    <organism ncbiTaxId="3702">
121      <names>
122        <shortLabel>arath</shortLabel>
123        <fullName>Arabidopsis thaliana (Mouse-ear cress)</fullName>
124      </names>
125    </organism>
126    <sequence>MILLQSHSRFLLQTLTRAQNLKAVELDYQWIEFDDVRYHVQVTMKNPNLLLSVSLPNPPPEAMSFGLPLGAEIAIKTTYGTGFQILDPPRDGFSLT
127      LKLNFSXVRPDELLTKLASIREVVMGAPLKIIFKHLASRTVAPELDRLVAIMHRPNETFFLVPQADKVTVAFPMPRFKDSVDTILATSFLKFEVEARRAA
128      ALMTAPSCSWSPAPQEQLEGAPKETLSANAGVFTVFIFPRHVGEKGLDRVTWNLSTFHAYVSYVHKFSEGFMHTRRRRRESMIQALDQAKPLEKTRSMN
129      NKSFRRLGLNEVNHNTNSK</sequence>
130  </interactor>
131  <interactor id="1441896">
132    <names>
133      <shortLabel>arpc4_arath</shortLabel>
134      <fullName>Actin-related protein 2/3 complex subunit 4</fullName>

```



```
135 <alias type="gene name synonym" typeAc="MI:0302">Actin-related protein C4</alias>
136 <alias type="gene name synonym" typeAc="MI:0302">Arp2/3 complex 20 kDa subunit</alias>
137 <alias type="orf name" typeAc="MI:0306">FCAALL.159</alias>
138 <alias type="orf name" typeAc="MI:0306">d13115c</alias>
139 <alias type="locus name" typeAc="MI:0305">At4g14147</alias>
140 <alias type="gene name" typeAc="MI:0301">ARPC4</alias>
141 </names>
142 <xref>
143 <primaryRef refTypeAc="MI:0356" refType="identity" id="F4JUL9" dbAc="MI:0486" db="uniprotkb"/>
144 <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-1547718" dbAc="MI:0469" db="intact"/>
145 <secondaryRef id="GO:0034314" dbAc="MI:0448" db="go"/>
146 <secondaryRef id="GO:0030041" dbAc="MI:0448" db="go"/>
147 <secondaryRef refTypeAc="MI:0360" refType="secondary-ac" version="SP_18" id="023274" dbAc="MI:0486" db="uniprotkb"/>
148 <secondaryRef id="GO:0005737" dbAc="MI:0448" db="go"/>
149 <secondaryRef id="IPRO08384" dbAc="MI:0449" db="interpro"/>
150 <secondaryRef id="GO:0005885" dbAc="MI:0448" db="go"/>
151 <secondaryRef id="GO:0042995" dbAc="MI:0448" db="go"/>
152 <secondaryRef id="NP_001031632.1" dbAc="MI:0481" db="refseq"/>
153 <secondaryRef id="IPI00656811" dbAc="MI:0675" db="ipi"/>
154 </xref>
155 <interactorType>
156 <names>
157 <shortLabel>protein</shortLabel>
158 <fullName>protein</fullName>
159 </names>
160 <xref>
161 <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0326" dbAc="MI:0488" db="psi-mi"/>
162 <secondaryRef refTypeAc="MI:0361" refType="see-also" id="SO:0000358" dbAc="MI:0601" db="so"/>
163 <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-619654" dbAc="MI:0469" db="intact"/>
164 <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
165 </xref>
166 </interactorType>
167 <organism ncbiTaxId="3702">
168 <names>
169 <shortLabel>arath</shortLabel>
170 <fullName>Arabidopsis thaliana (Mouse-ear cress)</fullName>
171 </names>
172 </organism>
173 <sequence>MANSRLRLYLACIKNTLEAAMCLQNFPCQEVERHNKPEVELKTSPELLLNPLVICRNEAEKCLIETSINSLRISLKVQADELENILTKKFLRFLSMRAEA
174 FQVLRKRPVQQYDISFLITNYHCEEMQKQLIDFIIQFMEDIEKEIRDLESVNTGRGLVATEFLKQFM</sequence>
175 </interactor>
176 </interactorList>
177 <interactionList>
178 <interaction id="1441897" imexId="IM-18781-1">
179 <names>
180 <shortLabel>arpc4-arpc2a</shortLabel>
181 </names>
182 <xref>
183 <primaryRef refTypeAc="MI:0356" refType="identity" id="EBI-6620113" dbAc="MI:0469" db="intact"/>
184 <secondaryRef refTypeAc="MI:0662" refType="imex-primary" id="IM-18781-1" dbAc="MI:0670" db="imex"/>
185 </xref>
186 <experimentList>
187 <experimentRef>1441894</experimentRef>
188 </experimentList>
189 <participantList>
190 <participant id="1441898">
191 <names>
192 <shortLabel>n/a</shortLabel>
193 </names>
194 <xref>
195 <primaryRef refTypeAc="MI:0356" refType="identity" id="EBI-6620116" dbAc="MI:0469" db="intact"/>
196 </xref>
197 <interactorRef>1441895</interactorRef>
198 <biologicalRole>
199 <names>
200 <shortLabel>unspecified role</shortLabel>
201 <fullName>unspecified role</fullName>
202 </names>
203 <xref>
204 <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0499" dbAc="MI:0488" db="psi-mi"/>
205 <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-77781" dbAc="MI:0469" db="intact"/>
206 <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
207 </xref>
208 </biologicalRole>
209 <experimentalRoleList>
210 <experimentalRole>
```

```

211     <names>
212         <shortLabel>prey</shortLabel>
213         <fullName>prey</fullName>
214     </names>
215     <xref>
216         <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0498" dbAc="MI:0488" db="psi-mi"/>
217         <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-58" dbAc="MI:0469" db="intact"/>
218         <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
219     </xref>
220 </experimentalRole>
221 </experimentalRoleList>
222 </participant>
223 <participant id="1441899">
224     <names>
225         <shortLabel>n/a</shortLabel>
226     </names>
227     <xref>
228         <primaryRef refTypeAc="MI:0356" refType="identity" id="EBI-6620115" dbAc="MI:0469" db="intact"/>
229     </xref>
230 <interactorRef>1441896</interactorRef>
231 <biologicalRole>
232     <names>
233         <shortLabel>unspecified role</shortLabel>
234         <fullName>unspecified role</fullName>
235     </names>
236     <xref>
237         <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0499" dbAc="MI:0488" db="psi-mi"/>
238         <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-77781" dbAc="MI:0469" db="intact"/>
239         <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
240     </xref>
241 </biologicalRole>
242 <experimentalRoleList>
243     <experimentalRole>
244         <names>
245             <shortLabel>bait</shortLabel>
246             <fullName>bait</fullName>
247         </names>
248         <xref>
249             <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0496" dbAc="MI:0488" db="psi-mi"/>
250             <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-49" dbAc="MI:0469" db="intact"/>
251             <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
252         </xref>
253     </experimentalRole>
254 </experimentalRoleList>
255 </participant>
256 </participantList>
257 <interactionType>
258     <names>
259         <shortLabel>physical association</shortLabel>
260         <fullName>physical association</fullName>
261     </names>
262     <xref>
263         <primaryRef refTypeAc="MI:0356" refType="identity" id="MI:0915" dbAc="MI:0488" db="psi-mi"/>
264         <secondaryRef refTypeAc="MI:0356" refType="identity" id="EBI-1813147" dbAc="MI:0469" db="intact"/>
265         <secondaryRef refTypeAc="MI:0358" refType="primary-reference" id="14755292" dbAc="MI:0446" db="pubmed"/>
266     </xref>
267 </interactionType>
268 <modelled>>false</modelled>
269 <intraMolecular>>false</intraMolecular>
270 <negative>>false</negative>
271 <attributeList>
272     <attribute nameAc="MI:0599" name="figure legend">Fig. 4b</attribute>
273 </attributeList>
274 </interaction>
275 </interactionList>
276 </entry>
277 </entrySet>

```

Listing 1: HUPO-PSI-MI XML 2.5 representation of a binary PPI

```

1 #ID(s) interactor A ID(s) interactor B Alt. ID(s) interactor A Alt. ID(s) interactor B Alias(es) interactor A Alias(es) interactor
  B Interaction detection method(s) Publication 1st author(s) Publication Identifier(s) Taxid interactor A Taxid interactor B
  Interaction type(s) Source database(s) Interaction identifier(s) Confidence value(s)
2 uniprotkb:Q8LGI3 uniprotkb:F4JUL9 intact:EBI-1547736|uniprotkb:Q9SY27 intact:EBI-1547718|uniprotkb:023274 psi-mi:arc2a_arath(
  display_long)|uniprotkb:ARPC2A(gene name)|psi-mi:ARPC2A(display_short)|uniprotkb:Actin-related protein C2A(gene name synonym)|
  uniprotkb:Arp2/3 complex 34 kDa subunit(gene name synonym)|uniprotkb:DIS2(gene name synonym)|uniprotkb:Protein DISTORTED
  TRICHOMES 2(gene name synonym)|uniprotkb:T17H7.13(orf name)|uniprotkb:At1g30825(locus name) psi-mi:arpc4_arath(display_long)|
  uniprotkb:Actin-related protein C4(gene name synonym)|uniprotkb:Arp2/3 complex 20 kDa subunit(gene name synonym)|uniprotkb:
  FCAALL.159(orf name)|uniprotkb:dl3115c(orf name)|uniprotkb:At4g14147(locus name)|uniprotkb:ARPC4(gene name)|psi-mi:ARPC4(
  display_short) psi-mi:"MI:0018"(two hybrid) El-Din et al. (2004) imex:IM-18781|pubmed:15086808 taxid:3702(arath)|taxid:3702("
  Arabidopsis thaliana (Mouse-ear cress)") taxid:3702(arath)|taxid:3702("Arabidopsis thaliana (Mouse-ear cress)") psi-mi:"MI
  :0915"(physical association) psi-mi:"MI:0486"(UniProt) intact:EBI-6620113|imex:IM-18781-1 intact-miscore:0.62

```

Listing 2: HUPO-PSI-MITAB 2.5 representation of a binary PPI

```

1 #ID(s) interactor A ID(s) interactor B Alt. ID(s) interactor A Alt. ID(s) interactor B Alias(es) interactor A Alias(es) interactor
  B Interaction detection method(s) Publication 1st author(s) Publication Identifier(s) Taxid interactor A Taxid interactor B
  Interaction type(s) Source database(s) Interaction identifier(s) Confidence value(s) Expansion method(s) Biological role(s)
  interactor A Biological role(s) interactor B Experimental role(s) interactor A Experimental role(s) interactor B Type(s)
  interactor A Type(s) interactor B Xref(s) interactor A Xref(s) interactor B Interaction Xref(s) Annotation(s) interactor A
  Annotation(s) interactor B Interaction annotation(s) Host organism(s) Interaction parameter(s) Creation date Update date
  Checksum(s) interactor A Checksum(s) interactor B Interaction Checksum(s) Negative Feature(s) interactor A Feature(s)
  interactor B Stoichiometry(s) interactor A Stoichiometry(s) interactor B Identification method participant A Identification
  method participant B
2 uniprotkb:Q8LGI3 uniprotkb:F4JUL9 intact:EBI-1547736|uniprotkb:Q9SY27 intact:EBI-1547718|uniprotkb:023274 psi-mi:arc2a_arath(
  display_long)|uniprotkb:ARPC2A(gene name)|psi-mi:ARPC2A(display_short)|uniprotkb:Actin-related protein C2A(gene name synonym)|
  uniprotkb:Arp2/3 complex 34 kDa subunit(gene name synonym)|uniprotkb:DIS2(gene name synonym)|uniprotkb:Protein DISTORTED
  TRICHOMES 2(gene name synonym)|uniprotkb:T17H7.13(orf name)|uniprotkb:At1g30825(locus name) psi-mi:arpc4_arath(display_long)|
  uniprotkb:Actin-related protein C4(gene name synonym)|uniprotkb:Arp2/3 complex 20 kDa subunit(gene name synonym)|uniprotkb:
  FCAALL.159(orf name)|uniprotkb:dl3115c(orf name)|uniprotkb:At4g14147(locus name)|uniprotkb:ARPC4(gene name)|psi-mi:ARPC4(
  display_short) psi-mi:"MI:0018"(two hybrid) El-Din et al. (2004) imex:IM-18781|pubmed:15086808 taxid:3702(arath)|taxid:3702("
  Arabidopsis thaliana (Mouse-ear cress)") taxid:3702(arath)|taxid:3702("Arabidopsis thaliana (Mouse-ear cress)") psi-mi:"MI
  :0915"(physical association) psi-mi:"MI:0486"(UniProt) intact:EBI-6620113|imex:IM-18781-1 intact-miscore:0.62 - psi-mi:"MI
  :0499"(unspecified role) psi-mi:"MI:0499"(unspecified role) psi-mi:"MI:0498"(prey) psi-mi:"MI:0496"(bait) psi-mi:"MI:0326"(
  protein) psi-mi:"MI:0326"(protein) go:"GO:0005885"(Arp2/3 protein complex)|go:"GO:0007015"(actin filament organization)|go:"GO
  :0030833"(regulation of actin filament polymerization)|go:"GO:0010090"(trichome morphogenesis)|refseq:NP_564364.1|interpro:
  IPR007188(Arp2/3 complex, 34kDa subunit p34-Arc)|ipi:IPI00521701|go:"GO:0005737"(cytoplasm)|go:"GO:0042995"(cell projection)|
  ipi:IPI00521322 go:"GO:0005885"(Arp2/3 protein complex)|go:"GO:0034314"(Arp2/3 complex-mediated actin nucleation)|refseq:
  NP_001031632.1|interpro:IPR008384(ARP23 complex 20 kDa subunit)|go:"GO:0030041"(actin filament polymerization)|ipi:IPI00656811|
  go:"GO:0005737"(cytoplasm)|go:"GO:0042995"(cell projection) - - - figure legend:Fig. 4b|curation depth:imex curation|full
  coverage:Only protein-protein interactions taxid:4932(yeast)|taxid:4932("Saccharomyces cerevisiae (Baker's yeast)") -
  2013/04/15 2013/04/15 rogid:3emb7Lf+rsEFn3GRcWEVJ3h4juM3702 rogid:0AxfqY3DADMOUCAGV4KFZE0AG1E3702 intact-crc:FE206A152B1CDEF2
  |rigid:zv0hbWzadLUgqOy1bHd4otPg0 false - - - psi-mi:"MI:0396"(predetermined participant) psi-mi:"MI:0396"(predetermined
  participant)

```

Listing 3: HUPO-PSI-MITAB 2.7 representation of a binary PPI

DASINT XML Schema Definition

The following listing contains the XML Schema Definition (XSD) of the DASINT XML format, which defines the response format of a DASMI server.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns="http://dasmi.de/" targetNamespace="http://dasmi.de/" elementFormDefault
   = "qualified" attributeFormDefault="unqualified">
3 <xs:element name="DASINT">
4 <xs:annotation>
5 <xs:documentation>The root element, does not contain any particular information. </xs:documentation>
6 </xs:annotation>
7 <xs:complexType>
8 <xs:sequence>
9 <xs:element name="INTERACTOR" maxOccurs="unbounded">
10 <xs:annotation>
11 <xs:documentation>The INTERACTOR element contains general information on the molecules participating in the interactions. The
   reference attributes like dbSource, dbAccessionId and dbVersion enable an unambiguous identification of the molecule. </xs:
   documentation>
12 </xs:annotation>
13 <xs:complexType>
14 <xs:sequence minOccurs="0">
15 <xs:element name="DETAIL" minOccurs="0" maxOccurs="unbounded">
16 <xs:annotation>
17 <xs:documentation>Additional information on theinteractor, e.g., synonyms or functional annotations.</xs:documentation>
18 </xs:annotation>
19 <xs:complexType>
20 <xs:sequence minOccurs="0">
21 <xs:element name="RANGE" minOccurs="0" maxOccurs="unbounded">
22 <xs:annotation>
23 <xs:documentation>Indicating that the father DETAIL element is a positional detail. For example, as a child of an INTERACTOR detail
   it can specify domains or strands on the interactor sequence.</xs:documentation>
24 </xs:annotation>
25 <xs:complexType>
26 <xs:attribute name="start" type="xs:int" use="required">
27 <xs:annotation>
28 <xs:documentation>The starting position of the detail in the sequence, e.g., 42.</xs:documentation>
29 </xs:annotation>
30 </xs:attribute>
31 <xs:attribute name="startStatus" type="xs:string" use="optional">
32 <xs:annotation>
33 <xs:documentation>The status of the starting position, e.g., certain.</xs:documentation>
34 </xs:annotation>
35 </xs:attribute>
36 <xs:attribute name="startStatusCvId" type="xs:string" use="optional">
37 <xs:annotation>
38 <xs:documentation> The controlled vocabulary term for the status of the starting position, e.g., MI:0335.</xs:documentation>
39 </xs:annotation>
40 </xs:attribute>
41 <xs:attribute name="end" type="xs:int" use="required">
42 <xs:annotation>
43 <xs:documentation>The ending position of the detail in the sequence, e.g., 91.</xs:documentation>
44 </xs:annotation>
45 </xs:attribute>
46 <xs:attribute name="endStatus" type="xs:string" use="optional">
47 <xs:annotation>
48 <xs:documentation>The status of the ending position, e.g., less-than.</xs:documentation>
49 </xs:annotation>
50 </xs:attribute>
51 <xs:attribute name="endStatusCvId" type="xs:string" use="optional">
52 <xs:annotation>
53 <xs:documentation>The controlled vocabulary term for the status of the ending position, e.g., MI:0337.</xs:documentation>
54 </xs:annotation>
55 </xs:attribute>
56 </xs:complexType>
57 </xs:element>
58 </xs:sequence>
59 <xs:attribute name="property" type="xs:string" use="required">
60 <xs:annotation>
61 <xs:documentation>The type of information the DETAIL element contains, e.g., experimentalRole, hostOrganism or authorConfidence.</xs:
   documentation>
62 </xs:annotation>
63 </xs:attribute>
64 <xs:attribute name="propertyCvId" type="xs:string" use="optional">
65 <xs:annotation>

```

```
66 | <xs:documentation>The controlled vocabulary term for the property, e.g., MI:0346. </xs:documentation>
67 | </xs:annotation>
68 | </xs:attribute>
69 | <xs:attribute name="value" type="xs:string" use="required">
70 | <xs:annotation>
71 | <xs:documentation>The actual piece of information, e.g., prey, human, or 1.0.</xs:documentation>
72 | </xs:annotation>
73 | </xs:attribute>
74 | <xs:attribute name="valueCvId" type="xs:string" use="optional">
75 | <xs:annotation>
76 | <xs:documentation>The controlled vocabulary term for the value, e.g., MI:0495. </xs:documentation>
77 | </xs:annotation>
78 | </xs:attribute>
79 | </xs:complexType>
80 | </xs:element>
81 | <xs:element name="SEQUENCE" minOccurs="0">
82 | <xs:annotation>
83 | <xs:documentation>The sequence of the interactor. The coordinate system of the interactor defines whether the sequence contains
    | peptides, nucleotides, or atoms.</xs:documentation>
84 | </xs:annotation>
85 | <xs:complexType>
86 | <xs:simpleContent>
87 | <xs:extension base="xs:string">
88 | <xs:attribute name="start" type="xs:string" use="optional">
89 | <xs:annotation>
90 | <xs:documentation>The starting position of the sequence.</xs:documentation>
91 | </xs:annotation>
92 | </xs:attribute>
93 | <xs:attribute name="end" type="xs:string" use="optional">
94 | <xs:annotation>
95 | <xs:documentation>The ending position of the sequence.</xs:documentation>
96 | </xs:annotation>
97 | </xs:attribute>
98 | </xs:extension>
99 | </xs:simpleContent>
100 | </xs:complexType>
101 | </xs:element>
102 | </xs:sequence>
103 | <xs:attribute name="intId" type="xs:string" use="required">
104 | <xs:annotation>
105 | <xs:documentation>Internal identifier that is used for cross-referencing between interactors and participants.</xs:documentation>
106 | </xs:annotation>
107 | </xs:attribute>
108 | <xs:attribute name="shortLabel" type="xs:string" use="required">
109 | <xs:annotation>
110 | <xs:documentation>The short name of the interacting molecule, e.g., hd_human or 14-3-3. Further names, like the full name or synonyms
    | can be stored in additional DETAIL elements. </xs:documentation>
111 | </xs:annotation>
112 | </xs:attribute>
113 | <xs:attribute name="dbSource" type="xs:string" use="required">
114 | <xs:annotation>
115 | <xs:documentation>The name of the primary referenced database, e.g., uniprotkb.</xs:documentation>
116 | </xs:annotation>
117 | </xs:attribute>
118 | <xs:attribute name="dbSourceCvId" type="xs:string" use="optional">
119 | <xs:annotation>
120 | <xs:documentation>The controlled vocabulary term for the primary database, e.g., MI:0486.</xs:documentation>
121 | </xs:annotation>
122 | </xs:attribute>
123 | <xs:attribute name="dbVersion" type="xs:string" use="optional">
124 | <xs:annotation>
125 | <xs:documentation>The version of the database, e.g., SP_26.</xs:documentation>
126 | </xs:annotation>
127 | </xs:attribute>
128 | <xs:attribute name="dbAccessionId" type="xs:string" use="required">
129 | <xs:annotation>
130 | <xs:documentation>The identifier of the reference in the primary database, e.g., P43246.</xs:documentation>
131 | </xs:annotation>
132 | </xs:attribute>
133 | <xs:attribute name="dbCoordSys" type="xs:string" use="optional">
134 | <xs:annotation>
135 | <xs:documentation>The coordinate system of the database, e.g., UniProt,Protein Sequence (see http://www.dasregistry.org/help\_coordsys.jsp). </xs:documentation>
136 | </xs:annotation>
137 | </xs:attribute>
138 | </xs:complexType>
139 | </xs:element>
```

```

140 <xs:element name="INTERACTION" maxOccurs="unbounded">
141 <xs:annotation>
142 <xs:documentation>The INTERACTION element contains information on a molecular interaction. An interaction is defined as an action
    between at least two objects. Accordingly, the INTERACTION element must contain at least two PARTICIPANT elements. </xs:
    documentation>
143 </xs:annotation>
144 <xs:complexType>
145 <xs:sequence>
146 <xs:element name="DETAIL" minOccurs="0" maxOccurs="unbounded">
147 <xs:annotation>
148 <xs:documentation>Additional information on the molecular interaction, e.g., a confidence score or information on the experiment.</xs
    :documentation>
149 </xs:annotation>
150 <xs:complexType>
151 <xs:attribute name="property" type="xs:string" use="required">
152 <xs:annotation>
153 <xs:documentation>The type of information the DETAIL element contains, e.g., experimentalRole, hostOrganism or authorConfidence. </xs
    :documentation>
154 </xs:annotation>
155 </xs:attribute>
156 <xs:attribute name="propertyCvId" type="xs:string" use="optional">
157 <xs:annotation>
158 <xs:documentation>The controlled vocabulary term for the property, e.g., MI:0346. </xs:documentation>
159 </xs:annotation>
160 </xs:attribute>
161 <xs:attribute name="value" type="xs:string" use="required">
162 <xs:annotation>
163 <xs:documentation>The actual piece of information, e.g., prey, human or 1,0.</xs:documentation>
164 </xs:annotation>
165 </xs:attribute>
166 <xs:attribute name="valueCvId" type="xs:string" use="optional">
167 <xs:annotation>
168 <xs:documentation>The controlled vocabulary term for the value, e.g., MI:0495. </xs:documentation>
169 </xs:annotation>
170 </xs:attribute>
171 </xs:complexType>
172 </xs:element>
173 <xs:element name="PARTICIPANT" minOccurs="2" maxOccurs="unbounded">
174 <xs:annotation>
175 <xs:documentation>The interactor participating in an interaction. The attribtue intId is used to link the PARTICIPANT to the initial
    INTERACTOR definition.</xs:documentation>
176 </xs:annotation>
177 <xs:complexType>
178 <xs:sequence minOccurs="0">
179 <xs:element name="DETAIL" minOccurs="0" maxOccurs="unbounded">
180 <xs:annotation>
181 <xs:documentation>Additional information on the participant, e.g., its experimental role.</xs:documentation>
182 </xs:annotation>
183 <xs:complexType>
184 <xs:sequence minOccurs="0">
185 <xs:element name="RANGE" minOccurs="0" maxOccurs="unbounded">
186 <xs:annotation>
187 <xs:documentation>Indicating that the father DETAIL element is a positional detail. For example, as a child of a PARTICIPANT DETAIL
    it can indicate a binding site. </xs:documentation>
188 </xs:annotation>
189 <xs:complexType>
190 <xs:attribute name="start" type="xs:int" use="required">
191 <xs:annotation>
192 <xs:documentation>The starting position of the detail in the sequence, e.g., 42.</xs:documentation>
193 </xs:annotation>
194 </xs:attribute>
195 <xs:attribute name="startStatus" type="xs:string" use="optional">
196 <xs:annotation>
197 <xs:documentation>The status of the starting position, e.g., certain.</xs:documentation>
198 </xs:annotation>
199 </xs:attribute>
200 <xs:attribute name="startStatusCvId" type="xs:string" use="optional">
201 <xs:annotation>
202 <xs:documentation> The controlled vocabulary term for the status of the starting position, e.g., MI:0335.</xs:documentation>
203 </xs:annotation>
204 </xs:attribute>
205 <xs:attribute name="end" type="xs:int" use="required">
206 <xs:annotation>
207 <xs:documentation>The ending position of the detail in the sequence, e.g., 91.</xs:documentation>
208 </xs:annotation>
209 </xs:attribute>
210 <xs:attribute name="endStatus" type="xs:string" use="optional">

```

```
211 <xs:annotation>
212 <xs:documentation>The status of the ending position, e.g., less-than.</xs:documentation>
213 </xs:annotation>
214 </xs:attribute>
215 <xs:attribute name="endStatusCvId" type="xs:string" use="optional">
216 <xs:annotation>
217 <xs:documentation>The controlled vocabulary term for the status of the ending position, e.g., MI:0337.</xs:documentation>
218 </xs:annotation>
219 </xs:attribute>
220 </xs:complexType>
221 </xs:element>
222 </xs:sequence>
223 <xs:attribute name="property" type="xs:string" use="required">
224 <xs:annotation>
225 <xs:documentation>The type of information the DETAIL element contains, e.g., experimentalRole, hostOrganism or authorConfidence.</xs:
documentation>
226 </xs:annotation>
227 </xs:attribute>
228 <xs:attribute name="propertyCvId" type="xs:string" use="optional">
229 <xs:annotation>
230 <xs:documentation>The controlled vocabulary term for the property, e.g., MI:0346. </xs:documentation>
231 </xs:annotation>
232 </xs:attribute>
233 <xs:attribute name="value" type="xs:string" use="required">
234 <xs:annotation>
235 <xs:documentation>The actual piece of information, e.g., prey, human or 1.0.</xs:documentation>
236 </xs:annotation>
237 </xs:attribute>
238 <xs:attribute name="valueCvId" type="xs:string" use="optional">
239 <xs:annotation>
240 <xs:documentation>The controlled vocabulary term for the value, e.g., MI:0495.</xs:documentation>
241 </xs:annotation>
242 </xs:attribute>
243 </xs:complexType>
244 </xs:element>
245 </xs:sequence>
246 <xs:attribute name="intId" type="xs:string" use="required">
247 <xs:annotation>
248 <xs:documentation>Internal identifier used to link an interactor participating in an interaction (-&gt; PARTICIPANT) to its
definition outside of the interaction (-&gt; INTERACTOR).</xs:documentation>
249 </xs:annotation>
250 </xs:attribute>
251 </xs:complexType>
252 </xs:element>
253 </xs:sequence>
254 <xs:attribute name="name" type="xs:string" use="required">
255 <xs:annotation>
256 <xs:documentation>The name of the interaction, e.g. exo1-msh2.</xs:documentation>
257 </xs:annotation>
258 </xs:attribute>
259 <xs:attribute name="dbSource" type="xs:string" use="required">
260 <xs:annotation>
261 <xs:documentation>The name of the referenced database, e.g., mint.</xs:documentation>
262 </xs:annotation>
263 </xs:attribute>
264 <xs:attribute name="dbSourceCvId" type="xs:string" use="optional">
265 <xs:annotation>
266 <xs:documentation>The controlled vocabulary term for the database, e.g., MI:0471.</xs:documentation>
267 </xs:annotation>
268 </xs:attribute>
269 <xs:attribute name="dbVersion" type="xs:string" use="optional">
270 <xs:annotation>
271 <xs:documentation>The version of the database, e.g., SP_26.</xs:documentation>
272 </xs:annotation>
273 </xs:attribute>
274 <xs:attribute name="dbAccessionId" type="xs:string" use="required">
275 <xs:annotation>
276 <xs:documentation>The identifier of the reference in the external database, e.g., MINT-84789.</xs:documentation>
277 </xs:annotation>
278 </xs:attribute>
279 </xs:complexType>
280 </xs:element>
281 </xs:sequence>
282 </xs:complexType>
283 </xs:element>
284 </xs:schema>
```

Listing 4: DASINT XML Schema Definition

PSISCORE Web Service Description Language

The following listing contains the PSISCORE Web Service Description Language (WSDL) that defines how PSISCORE servers can be accessed programmatically.

```

1  <?xml version='1.0' encoding='UTF-8'?>
2  <wsdl:definitions name="PsiscoreService" targetNamespace="http://psi.hupo.org/mi/psiscore" xmlns:ns1="http://cxf.apache.org/bindings/
   xformat" xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" xmlns:tns="http://psi.hupo.org/mi/psiscore" xmlns:wsdl="http://
   schemas.xmlsoap.org/wsdl/" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
3  <wsdl:types>
4  <xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" targetNamespace="http://psi.hupo.org/mi/psiscore" xmlns:
   ns1="http://psi.hupo.org/mi/mif" xmlns:tns="http://psi.hupo.org/mi/psiscore" xmlns:xs="http://www.w3.org/2001/XMLSchema">
5  <xs:import namespace="http://psi.hupo.org/mi/mif" />
6  <xs:element name="getJob" type="tns:getJob" />
7  <xs:element name="getJobResponse" type="tns:getJobResponse" />
8  <xs:element name="getJobStatus" type="tns:getJobStatus" />
9  <xs:element name="getJobStatusResponse" type="tns:getJobStatusResponse" />
10 <xs:element name="getSupportedDataTypes" type="tns:getSupportedDataTypes" />
11 <xs:element name="getSupportedDataTypesResponse" type="tns:getSupportedDataTypesResponse" />
12 <xs:element name="getSupportedDbAcs" type="tns:getSupportedDbAcs" />
13 <xs:element name="getSupportedDbAcsResponse" type="tns:getSupportedDbAcsResponse" />
14 <xs:element name="getSupportedScoringMethods" type="tns:getSupportedScoringMethods" />
15 <xs:element name="getSupportedScoringMethodsResponse" type="tns:getSupportedScoringMethodsResponse" />
16 <xs:element name="getVersion" type="tns:getVersion" />
17 <xs:element name="getVersionResponse" type="tns:getVersionResponse" />
18 <xs:element name="submitJob" type="tns:submitJob" />
19 <xs:element name="submitJobResponse" type="tns:submitJobResponse" />
20 <xs:complexType name="getSupportedScoringMethods">
21 <xs:sequence />
22 </xs:complexType>
23 <xs:complexType name="getSupportedScoringMethodsResponse">
24 <xs:sequence>
25 <xs:element maxOccurs="unbounded" minOccurs="0" name="algorithmDescriptor" type="tns:algorithmDescriptor" />
26 </xs:sequence>
27 </xs:complexType>
28 <xs:complexType name="algorithmDescriptor">
29 <xs:sequence>
30 <xs:element name="id" type="xs:string" />
31 <xs:element maxOccurs="unbounded" minOccurs="0" name="algorithmType" type="xs:string" />
32 <xs:element maxOccurs="unbounded" minOccurs="0" name="parameterSpecifier" type="tns:parameterSpecifier" />
33 <xs:element maxOccurs="unbounded" minOccurs="0" name="requiredDataFields" nillable="true" type="xs:string" />
34 <xs:element minOccurs="0" name="range" type="xs:string" />
35 </xs:sequence>
36 </xs:complexType>
37 <xs:complexType name="parameterSpecifier">
38 <xs:sequence>
39 <xs:element name="id" type="xs:string" />
40 <xs:element name="type" type="xs:string" />
41 <xs:element name="value" nillable="true" type="xs:string" />
42 </xs:sequence>
43 </xs:complexType>
44 <xs:complexType name="psiscoreFault">
45 <xs:sequence>
46 <xs:element name="code" type="xs:int" />
47 <xs:element name="message" type="xs:string" />
48 </xs:sequence>
49 </xs:complexType>
50 <xs:complexType name="submitJob">
51 <xs:sequence>
52 <xs:element maxOccurs="unbounded" minOccurs="0" name="algorithmDescriptor" type="tns:algorithmDescriptor" />
53 <xs:element name="inputData" type="tns:resultSet" />
54 <xs:element name="returnFormat" type="xs:string" />
55 </xs:sequence>
56 </xs:complexType>
57 <xs:complexType name="resultSet">
58 <xs:sequence>
59 <xs:element minOccurs="0" name="entrySet" type="ns1:entrySet" />
60 <xs:element minOccurs="0" name="mitab" type="xs:string" />
61 </xs:sequence>
62 </xs:complexType>
63 <xs:complexType name="submitJobResponse">
64 <xs:sequence>
65 <xs:element name="jobResponse" type="tns:jobResponse" />
66 </xs:sequence>
67 </xs:complexType>

```

```
68 <xs:complexType name="jobResponse">
69 <xs:sequence>
70 <xs:element name="jobId" type="xs:string" />
71 <xs:element name="pollingInterval" type="xs:int" />
72 </xs:sequence>
73 </xs:complexType>
74 <xs:complexType name="getSupportedDataTypes">
75 <xs:sequence />
76 </xs:complexType>
77 <xs:complexType name="getSupportedDataTypesResponse">
78 <xs:sequence>
79 <xs:element maxOccurs="unbounded" name="return" type="xs:string" />
80 </xs:sequence>
81 </xs:complexType>
82 <xs:complexType name="getVersion">
83 <xs:sequence />
84 </xs:complexType>
85 <xs:complexType name="getVersionResponse">
86 <xs:sequence>
87 <xs:element minOccurs="0" name="return" type="xs:string" />
88 </xs:sequence>
89 </xs:complexType>
90 <xs:complexType name="getJobStatus">
91 <xs:sequence>
92 <xs:element name="jobId" type="xs:string" />
93 </xs:sequence>
94 </xs:complexType>
95 <xs:complexType name="getJobStatusResponse">
96 <xs:sequence>
97 <xs:element name="jobStatus" type="xs:string" />
98 </xs:sequence>
99 </xs:complexType>
100 <xs:complexType name="getJob">
101 <xs:sequence>
102 <xs:element name="jobId" type="xs:string" />
103 </xs:sequence>
104 </xs:complexType>
105 <xs:complexType name="getJobResponse">
106 <xs:sequence>
107 <xs:element name="jobResponse" type="tns:queryResponse" />
108 </xs:sequence>
109 </xs:complexType>
110 <xs:complexType name="queryResponse">
111 <xs:sequence>
112 <xs:element name="resultSet" type="tns:resultSet" />
113 <xs:element name="report" type="tns:report" />
114 </xs:sequence>
115 </xs:complexType>
116 <xs:complexType name="report">
117 <xs:sequence>
118 <xs:element maxOccurs="unbounded" minOccurs="0" name="result" type="xs:string" />
119 </xs:sequence>
120 </xs:complexType>
121 <xs:complexType name="requestInfo">
122 <xs:sequence>
123 <xs:element name="resultType" type="xs:string" />
124 <xs:element default="0" name="firstResult" type="xs:int" />
125 <xs:element default="50" name="blockSize" type="xs:int" />
126 </xs:sequence>
127 </xs:complexType>
128 <xs:complexType name="getSupportedDbAcsResponse">
129 <xs:sequence>
130 <xs:element maxOccurs="unbounded" name="return" type="xs:string" />
131 </xs:sequence>
132 </xs:complexType>
133 <xs:complexType name="getSupportedDbAcs">
134 <xs:sequence />
135 </xs:complexType>
136 <xs:complexType name="supportedTypes">
137 <xs:sequence>
138 <xs:element maxOccurs="unbounded" name="supportedType" type="xs:string" />
139 </xs:sequence>
140 </xs:complexType>
141 <xs:complexType name="dbRef">
142 <xs:sequence>
143 <xs:element name="dbAc" nillable="true" type="xs:string" />
```

```
144 <xs:element name="id" type="xs:string" />
145 </xs:sequence>
146 </xs:complexType>
147 <xs:element name="PsiscoreException" nillable="true" type="tns:psiscoreFault" />
148 <xs:element name="InvalidArgumentException" nillable="true" type="tns:psiscoreFault" />
149 <xs:element name="JobStillRunningException" nillable="true" type="tns:psiscoreFault" />
150 </xs:schema>
151 <xs:schema elementFormDefault="qualified" targetNamespace="http://psi.hupo.org/mi/mif" version="1.0" xmlns:tns="http://psi.hupo.org/
mi/mif" xmlns:xs="http://www.w3.org/2001/XMLSchema">
152 <xs:element name="entrySet" type="tns:entrySet" />
153 <xs:complexType name="entrySet">
154 <xs:sequence>
155 <xs:element maxOccurs="unbounded" name="entry" type="tns:entry" />
156 </xs:sequence>
157 <xs:attribute name="level" type="xs:int" use="required" />
158 <xs:attribute name="version" type="xs:int" use="required" />
159 <xs:attribute name="minorVersion" type="xs:int" />
160 </xs:complexType>
161 <xs:complexType name="entry">
162 <xs:sequence>
163 <xs:element minOccurs="0" name="source" type="tns:source" />
164 <xs:element minOccurs="0" name="availabilityList" type="tns:availabilityList" />
165 <xs:element minOccurs="0" name="experimentList" type="tns:experimentDescriptionList" />
166 <xs:element minOccurs="0" name="interactorList" type="tns:interactorList" />
167 <xs:element name="interactionList" type="tns:interactionList" />
168 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
169 </xs:sequence>
170 </xs:complexType>
171 <xs:complexType name="source">
172 <xs:sequence>
173 <xs:element minOccurs="0" name="names" type="tns:names" />
174 <xs:element minOccurs="0" name="bibref" type="tns:bibref" />
175 <xs:element minOccurs="0" name="xref" type="tns:xref" />
176 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
177 </xs:sequence>
178 <xs:attribute name="release" type="xs:string" />
179 <xs:attribute name="releaseDate" type="xs:date" />
180 </xs:complexType>
181 <xs:complexType name="names">
182 <xs:sequence>
183 <xs:element minOccurs="0" name="shortLabel" type="xs:string" />
184 <xs:element minOccurs="0" name="fullName" type="xs:string" />
185 <xs:element maxOccurs="unbounded" minOccurs="0" name="alias" nillable="true" type="tns:alias" />
186 </xs:sequence>
187 </xs:complexType>
188 <xs:complexType name="alias">
189 <xs:simpleContent>
190 <xs:extension base="xs:string">
191 <xs:attribute name="typeAc" type="xs:string" />
192 <xs:attribute name="type" type="xs:string" />
193 </xs:extension>
194 </xs:simpleContent>
195 </xs:complexType>
196 <xs:complexType name="bibref">
197 <xs:sequence>
198 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
199 <xs:element minOccurs="0" name="xref" type="tns:xref" />
200 </xs:sequence>
201 </xs:complexType>
202 <xs:complexType name="attributeList">
203 <xs:sequence>
204 <xs:element maxOccurs="unbounded" name="attribute" type="tns:attribute" />
205 </xs:sequence>
206 </xs:complexType>
207 <xs:complexType name="attribute">
208 <xs:simpleContent>
209 <xs:extension base="xs:string">
210 <xs:attribute name="name" type="xs:string" use="required" />
211 <xs:attribute name="nameAc" type="xs:string" />
212 </xs:extension>
213 </xs:simpleContent>
214 </xs:complexType>
215 <xs:complexType name="xref">
216 <xs:sequence>
217 <xs:element name="primaryRef" type="tns:dbReference" />
218 <xs:element maxOccurs="unbounded" minOccurs="0" name="secondaryRef" type="tns:dbReference" />
```

```
219 </xs:sequence>
220 </xs:complexType>
221 <xs:complexType name="dbReference">
222 <xs:sequence>
223 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
224 </xs:sequence>
225 <xs:attribute name="db" type="xs:string" use="required" />
226 <xs:attribute name="dbAc" type="xs:string" />
227 <xs:attribute name="id" type="xs:string" use="required" />
228 <xs:attribute name="secondary" type="xs:string" />
229 <xs:attribute name="version" type="xs:string" />
230 <xs:attribute name="refType" type="xs:string" />
231 <xs:attribute name="refTypeAc" type="xs:string" />
232 </xs:complexType>
233 <xs:complexType name="availabilityList">
234 <xs:sequence>
235 <xs:element maxOccurs="unbounded" minOccurs="0" name="availability" type="tns:availability" />
236 </xs:sequence>
237 </xs:complexType>
238 <xs:complexType name="availability">
239 <xs:simpleContent>
240 <xs:extension base="xs:string">
241 <xs:attribute name="id" type="xs:int" use="required" />
242 </xs:extension>
243 </xs:simpleContent>
244 </xs:complexType>
245 <xs:complexType name="experimentDescriptionList">
246 <xs:sequence>
247 <xs:element maxOccurs="unbounded" minOccurs="0" name="experimentDescription" type="tns:experimentDescription" />
248 </xs:sequence>
249 </xs:complexType>
250 <xs:complexType name="experimentDescription">
251 <xs:sequence>
252 <xs:element minOccurs="0" name="names" type="tns:names" />
253 <xs:element name="bibref" type="tns:bibref" />
254 <xs:element minOccurs="0" name="xref" type="tns:xref" />
255 <xs:element minOccurs="0" name="hostOrganismList" type="tns:hostOrganismList" />
256 <xs:element name="interactionDetectionMethod" type="tns:cvType" />
257 <xs:element minOccurs="0" name="participantIdentificationMethod" type="tns:cvType" />
258 <xs:element minOccurs="0" name="featureDetectionMethod" type="tns:cvType" />
259 <xs:element minOccurs="0" name="confidenceList" type="tns:confidenceList" />
260 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
261 </xs:sequence>
262 <xs:attribute name="id" type="xs:int" use="required" />
263 </xs:complexType>
264 <xs:complexType name="hostOrganismList">
265 <xs:sequence>
266 <xs:element maxOccurs="unbounded" name="hostOrganism" type="tns:hostOrganism" />
267 </xs:sequence>
268 </xs:complexType>
269 <xs:complexType name="hostOrganism">
270 <xs:complexContent>
271 <xs:extension base="tns:bioSource">
272 <xs:sequence>
273 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
274 </xs:sequence>
275 </xs:extension>
276 </xs:complexContent>
277 </xs:complexType>
278 <xs:complexType name="bioSource">
279 <xs:sequence>
280 <xs:element minOccurs="0" name="names" type="tns:names" />
281 <xs:element minOccurs="0" name="cellType" type="tns:openCvType" />
282 <xs:element minOccurs="0" name="compartment" type="tns:openCvType" />
283 <xs:element minOccurs="0" name="tissue" type="tns:openCvType" />
284 </xs:sequence>
285 <xs:attribute name="ncbiTaxId" type="xs:int" use="required" />
286 </xs:complexType>
287 <xs:complexType name="experimentRefList">
288 <xs:sequence>
289 <xs:element maxOccurs="unbounded" minOccurs="0" name="experimentRef" type="xs:int" />
290 </xs:sequence>
291 </xs:complexType>
292 <xs:complexType name="openCvType">
293 <xs:sequence>
294 <xs:element name="names" type="tns:names" />
```

```
295 <xs:element minOccurs="0" name="xref" type="tns:xref" />
296 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
297 </xs:sequence>
298 </xs:complexType>
299 <xs:complexType name="cvType">
300 <xs:sequence>
301 <xs:element name="names" type="tns:names" />
302 <xs:element name="xref" type="tns:xref" />
303 </xs:sequence>
304 </xs:complexType>
305 <xs:complexType name="experimentalPreparation">
306 <xs:complexContent>
307 <xs:extension base="tns:cvType">
308 <xs:sequence>
309 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
310 </xs:sequence>
311 </xs:extension>
312 </xs:complexContent>
313 </xs:complexType>
314 <xs:complexType name="experimentalRole">
315 <xs:complexContent>
316 <xs:extension base="tns:cvType">
317 <xs:sequence>
318 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
319 </xs:sequence>
320 </xs:extension>
321 </xs:complexContent>
322 </xs:complexType>
323 <xs:complexType name="participantIdentificationMethod">
324 <xs:complexContent>
325 <xs:extension base="tns:cvType">
326 <xs:sequence>
327 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
328 </xs:sequence>
329 </xs:extension>
330 </xs:complexContent>
331 </xs:complexType>
332 <xs:complexType name="confidenceList">
333 <xs:sequence>
334 <xs:element maxOccurs="unbounded" name="confidence" type="tns:confidence" />
335 </xs:sequence>
336 </xs:complexType>
337 <xs:complexType name="confidence">
338 <xs:complexContent>
339 <xs:extension base="tns:confidenceBase">
340 <xs:sequence>
341 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
342 </xs:sequence>
343 </xs:extension>
344 </xs:complexContent>
345 </xs:complexType>
346 <xs:complexType name="confidenceBase">
347 <xs:sequence>
348 <xs:element name="unit" type="tns:openCvType" />
349 <xs:element name="value" type="xs:string" />
350 </xs:sequence>
351 </xs:complexType>
352 <xs:complexType name="interactorList">
353 <xs:sequence>
354 <xs:element maxOccurs="unbounded" minOccurs="0" name="interactor" type="tns:interactor" />
355 </xs:sequence>
356 </xs:complexType>
357 <xs:complexType name="interactor">
358 <xs:sequence>
359 <xs:element name="names" type="tns:names" />
360 <xs:element minOccurs="0" name="xref" type="tns:xref" />
361 <xs:element name="interactorType" type="tns:cvType" />
362 <xs:element minOccurs="0" name="organism">
363 <xs:complexType>
364 <xs:complexContent>
365 <xs:extension base="tns:bioSource">
366 <xs:sequence />
367 </xs:extension>
368 </xs:complexContent>
369 </xs:complexType>
370 </xs:element>
```

```

371 <xs:element minOccurs="0" name="sequence" type="xs:string" />
372 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
373 </xs:sequence>
374 <xs:attribute name="id" type="xs:int" use="required" />
375 </xs:complexType>
376 <xs:complexType name="interactionList">
377 <xs:sequence>
378 <xs:element maxOccurs="unbounded" name="interaction" type="tns:interaction" />
379 </xs:sequence>
380 </xs:complexType>
381 <xs:complexType name="interaction">
382 <xs:sequence>
383 <xs:element minOccurs="0" name="names" type="tns:names" />
384 <xs:element minOccurs="0" name="xref" type="tns:xref" />
385 <xs:element minOccurs="0" name="availability" type="tns:availability" />
386 <xs:element minOccurs="0" name="availabilityRef" type="xs:int" />
387 <xs:element minOccurs="0" name="experimentList" type="tns:experimentList" />
388 <xs:element name="participantList" type="tns:participantList" />
389 <xs:element minOccurs="0" name="inferredInteractionList" type="tns:inferredInteractionList" />
390 <xs:element maxOccurs="unbounded" minOccurs="0" name="interactionType" type="tns:cvType" />
391 <xs:element minOccurs="0" name="modelled" type="xs:boolean" />
392 <xs:element default="false" minOccurs="0" name="intraMolecular" type="xs:boolean" />
393 <xs:element default="false" minOccurs="0" name="negative" type="xs:boolean" />
394 <xs:element minOccurs="0" name="confidenceList" type="tns:confidenceList" />
395 <xs:element minOccurs="0" name="parameterList" type="tns:parameterList" />
396 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
397 </xs:sequence>
398 <xs:attribute name="imexId" type="xs:string" />
399 <xs:attribute name="id" type="xs:int" use="required" />
400 </xs:complexType>
401 <xs:complexType name="experimentList">
402 <xs:sequence>
403 <xs:choice maxOccurs="unbounded" minOccurs="0">
404 <xs:element name="experimentRef" type="xs:int" />
405 <xs:element name="experimentDescription" type="tns:experimentDescription" />
406 </xs:choice>
407 </xs:sequence>
408 </xs:complexType>
409 <xs:complexType name="participantList">
410 <xs:sequence>
411 <xs:element maxOccurs="unbounded" name="participant" type="tns:participant" />
412 </xs:sequence>
413 </xs:complexType>
414 <xs:complexType name="participant">
415 <xs:sequence>
416 <xs:element minOccurs="0" name="names" type="tns:names" />
417 <xs:element minOccurs="0" name="xref" type="tns:xref" />
418 <xs:element minOccurs="0" name="interactionRef" type="xs:int" />
419 <xs:element minOccurs="0" name="interactor" type="tns:interactor" />
420 <xs:element minOccurs="0" name="interactorRef" type="xs:int" />
421 <xs:element minOccurs="0" name="participantIdentificationMethodList" type="tns:participantIdentificationMethodList" />
422 <xs:element minOccurs="0" name="biologicalRole" type="tns:cvType" />
423 <xs:element minOccurs="0" name="experimentalRoleList" type="tns:experimentalRoleList" />
424 <xs:element minOccurs="0" name="experimentalPreparationList" type="tns:experimentalPreparationList" />
425 <xs:element minOccurs="0" name="experimentalInteractorList" type="tns:experimentalInteractorList" />
426 <xs:element minOccurs="0" name="featureList" type="tns:featureList" />
427 <xs:element minOccurs="0" name="hostOrganismList" type="tns:hostOrganismList" />
428 <xs:element minOccurs="0" name="confidenceList" type="tns:confidenceList" />
429 <xs:element minOccurs="0" name="parameterList" type="tns:parameterList" />
430 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
431 </xs:sequence>
432 <xs:attribute name="id" type="xs:int" use="required" />
433 </xs:complexType>
434 <xs:complexType name="participantIdentificationMethodList">
435 <xs:sequence>
436 <xs:element maxOccurs="unbounded" name="participantIdentificationMethod" type="tns:participantIdentificationMethod" />
437 </xs:sequence>
438 </xs:complexType>
439 <xs:complexType name="experimentalRoleList">
440 <xs:sequence>
441 <xs:element maxOccurs="unbounded" name="experimentalRole" type="tns:experimentalRole" />
442 </xs:sequence>
443 </xs:complexType>
444 <xs:complexType name="experimentalPreparationList">
445 <xs:sequence>
446 <xs:element maxOccurs="unbounded" name="experimentalPreparation" type="tns:experimentalPreparation" />

```

```
447 </xs:sequence>
448 </xs:complexType>
449 <xs:complexType name="experimentalInteractorList">
450 <xs:sequence>
451 <xs:element maxOccurs="unbounded" name="experimentalInteractor" type="tns:experimentalInteractor" />
452 </xs:sequence>
453 </xs:complexType>
454 <xs:complexType name="experimentalInteractor">
455 <xs:sequence>
456 <xs:element minOccurs="0" name="interactor" type="tns:interactor" />
457 <xs:element minOccurs="0" name="interactorRef" type="xs:int" />
458 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
459 </xs:sequence>
460 </xs:complexType>
461 <xs:complexType name="featureList">
462 <xs:sequence>
463 <xs:element maxOccurs="unbounded" name="feature" type="tns:feature" />
464 </xs:sequence>
465 </xs:complexType>
466 <xs:complexType name="feature">
467 <xs:sequence>
468 <xs:element minOccurs="0" name="names" type="tns:names" />
469 <xs:element minOccurs="0" name="xref" type="tns:xref" />
470 <xs:element minOccurs="0" name="featureType" type="tns:cvType" />
471 <xs:element minOccurs="0" name="featureDetectionMethod" type="tns:cvType" />
472 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
473 <xs:element name="featureRangeList">
474 <xs:complexType>
475 <xs:sequence>
476 <xs:element maxOccurs="unbounded" name="featureRange" type="tns:baseLocation" />
477 </xs:sequence>
478 </xs:complexType>
479 </xs:element>
480 <xs:element minOccurs="0" name="attributeList" type="tns:attributeList" />
481 </xs:sequence>
482 <xs:attribute name="id" type="xs:int" use="required" />
483 </xs:complexType>
484 <xs:complexType name="baseLocation">
485 <xs:sequence>
486 <xs:element name="startStatus" type="tns:cvType" />
487 <xs:element minOccurs="0" name="beginInterval" type="tns:interval" />
488 <xs:element minOccurs="0" name="begin" type="tns:position" />
489 <xs:element name="endStatus" type="tns:cvType" />
490 <xs:element minOccurs="0" name="endInterval" type="tns:interval" />
491 <xs:element minOccurs="0" name="end" type="tns:position" />
492 <xs:element default="false" minOccurs="0" name="isLink" type="xs:boolean" />
493 </xs:sequence>
494 </xs:complexType>
495 <xs:complexType name="interval">
496 <xs:sequence />
497 <xs:attribute name="begin" type="xs:unsignedLong" use="required" />
498 <xs:attribute name="end" type="xs:unsignedLong" use="required" />
499 </xs:complexType>
500 <xs:complexType name="position">
501 <xs:sequence />
502 <xs:attribute name="position" type="xs:unsignedLong" use="required" />
503 </xs:complexType>
504 <xs:complexType name="parameterList">
505 <xs:sequence>
506 <xs:element maxOccurs="unbounded" name="parameter" type="tns:parameter" />
507 </xs:sequence>
508 </xs:complexType>
509 <xs:complexType name="parameter">
510 <xs:complexContent>
511 <xs:extension base="tns:parameterBase">
512 <xs:sequence>
513 <xs:element name="experimentRef" type="xs:int" />
514 </xs:sequence>
515 <xs:attribute name="uncertainty" type="xs:decimal" />
516 </xs:extension>
517 </xs:complexContent>
518 </xs:complexType>
519 <xs:complexType name="parameterBase">
520 <xs:sequence />
521 <xs:attribute name="term" type="xs:string" use="required" />
522 <xs:attribute name="termAc" type="xs:string" />
```

```
523 <xs:attribute name="unit" type="xs:string" />
524 <xs:attribute name="unitAc" type="xs:string" />
525 <xs:attribute name="base" type="xs:short" />
526 <xs:attribute name="exponent" type="xs:short" />
527 <xs:attribute name="factor" type="xs:decimal" use="required" />
528 </xs:complexType>
529 <xs:complexType name="inferredInteractionList">
530 <xs:sequence>
531 <xs:element maxOccurs="unbounded" name="inferredInteraction" type="tns:inferredInteraction" />
532 </xs:sequence>
533 </xs:complexType>
534 <xs:complexType name="inferredInteraction">
535 <xs:sequence>
536 <xs:element maxOccurs="unbounded" name="participant" type="tns:inferredInteractionParticipant" />
537 <xs:element minOccurs="0" name="experimentRefList" type="tns:experimentRefList" />
538 </xs:sequence>
539 </xs:complexType>
540 <xs:complexType name="inferredInteractionParticipant">
541 <xs:sequence>
542 <xs:element minOccurs="0" name="participantFeatureRef" type="xs:int" />
543 <xs:element minOccurs="0" name="participantRef" type="xs:int" />
544 </xs:sequence>
545 </xs:complexType>
546 <xs:simpleType name="fullName">
547 <xs:restriction base="xs:string" />
548 </xs:simpleType>
549 <xs:simpleType name="label">
550 <xs:restriction base="xs:string" />
551 </xs:simpleType>
552 </xs:schema>
553 </wsdl:types>
554 <wsdl:message name="getJobResponse">
555 <wsdl:part element="tns:getJobResponse" name="parameters">
556 </wsdl:part>
557 </wsdl:message>
558 <wsdl:message name="getVersion">
559 <wsdl:part element="tns:getVersion" name="parameters">
560 </wsdl:part>
561 </wsdl:message>
562 <wsdl:message name="getJob">
563 <wsdl:part element="tns:getJob" name="parameters">
564 </wsdl:part>
565 </wsdl:message>
566 <wsdl:message name="PsiscoreException">
567 <wsdl:part element="tns:PsiscoreException" name="PsiscoreException">
568 </wsdl:part>
569 </wsdl:message>
570 <wsdl:message name="getSupportedDataTypes">
571 <wsdl:part element="tns:getSupportedDataTypes" name="parameters">
572 </wsdl:part>
573 </wsdl:message>
574 <wsdl:message name="InvalidArgumentException">
575 <wsdl:part element="tns:InvalidArgumentException" name="InvalidArgumentException">
576 </wsdl:part>
577 </wsdl:message>
578 <wsdl:message name="getJobStatus">
579 <wsdl:part element="tns:getJobStatus" name="parameters">
580 </wsdl:part>
581 </wsdl:message>
582 <wsdl:message name="getSupportedScoringMethodsResponse">
583 <wsdl:part element="tns:getSupportedScoringMethodsResponse" name="parameters">
584 </wsdl:part>
585 </wsdl:message>
586 <wsdl:message name="getJobStatusResponse">
587 <wsdl:part element="tns:getJobStatusResponse" name="parameters">
588 </wsdl:part>
589 </wsdl:message>
590 <wsdl:message name="getSupportedScoringMethods">
591 <wsdl:part element="tns:getSupportedScoringMethods" name="parameters">
592 </wsdl:part>
593 </wsdl:message>
594 <wsdl:message name="submitJob">
595 <wsdl:part element="tns:submitJob" name="parameters">
596 </wsdl:part>
597 </wsdl:message>
598 <wsdl:message name="getSupportedDataTypesResponse">
```



```
599 <wsdl:part element="tns:getSupportedDataTypesResponse" name="parameters">
600 </wsdl:part>
601 </wsdl:message>
602 <wsdl:message name="getVersionResponse">
603 <wsdl:part element="tns:getVersionResponse" name="parameters">
604 </wsdl:part>
605 </wsdl:message>
606 <wsdl:message name="JobStillRunningException">
607 <wsdl:part element="tns:JobStillRunningException" name="JobStillRunningException">
608 </wsdl:part>
609 </wsdl:message>
610 <wsdl:message name="submitJobResponse">
611 <wsdl:part element="tns:submitJobResponse" name="parameters">
612 </wsdl:part>
613 </wsdl:message>
614 <wsdl:portType name="psiscoreService">
615 <wsdl:operation name="getSupportedScoringMethods">
616 <wsdl:input message="tns:getSupportedScoringMethods" name="getSupportedScoringMethods">
617 </wsdl:input>
618 <wsdl:output message="tns:getSupportedScoringMethodsResponse" name="getSupportedScoringMethodsResponse">
619 </wsdl:output>
620 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
621 </wsdl:fault>
622 </wsdl:operation>
623 <wsdl:operation name="submitJob">
624 <wsdl:input message="tns:submitJob" name="submitJob">
625 </wsdl:input>
626 <wsdl:output message="tns:submitJobResponse" name="submitJobResponse">
627 </wsdl:output>
628 <wsdl:fault message="tns:InvalidArgumentException" name="InvalidArgumentException">
629 </wsdl:fault>
630 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
631 </wsdl:fault>
632 </wsdl:operation>
633 <wsdl:operation name="getSupportedDataTypes">
634 <wsdl:input message="tns:getSupportedDataTypes" name="getSupportedDataTypes">
635 </wsdl:input>
636 <wsdl:output message="tns:getSupportedDataTypesResponse" name="getSupportedDataTypesResponse">
637 </wsdl:output>
638 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
639 </wsdl:fault>
640 </wsdl:operation>
641 <wsdl:operation name="getVersion">
642 <wsdl:input message="tns:getVersion" name="getVersion">
643 </wsdl:input>
644 <wsdl:output message="tns:getVersionResponse" name="getVersionResponse">
645 </wsdl:output>
646 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
647 </wsdl:fault>
648 </wsdl:operation>
649 <wsdl:operation name="getJobStatus">
650 <wsdl:input message="tns:getJobStatus" name="getJobStatus">
651 </wsdl:input>
652 <wsdl:output message="tns:getJobStatusResponse" name="getJobStatusResponse">
653 </wsdl:output>
654 <wsdl:fault message="tns:InvalidArgumentException" name="InvalidArgumentException">
655 </wsdl:fault>
656 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
657 </wsdl:fault>
658 </wsdl:operation>
659 <wsdl:operation name="getJob">
660 <wsdl:input message="tns:getJob" name="getJob">
661 </wsdl:input>
662 <wsdl:output message="tns:getJobResponse" name="getJobResponse">
663 </wsdl:output>
664 <wsdl:fault message="tns:JobStillRunningException" name="JobStillRunningException">
665 </wsdl:fault>
666 <wsdl:fault message="tns:InvalidArgumentException" name="InvalidArgumentException">
667 </wsdl:fault>
668 <wsdl:fault message="tns:PsiscoreException" name="PsiscoreException">
669 </wsdl:fault>
670 </wsdl:operation>
671 </wsdl:portType>
672 <wsdl:binding name="PsiscoreServiceSoapBinding" type="tns:psiscoreService">
673 <soap:binding style="document" transport="http://schemas.xmlsoap.org/soap/http" />
674 <wsdl:operation name="getSupportedScoringMethods">
```

```
675 <soap:operation soapAction="getSupportedScoringMethods" style="document" />
676 <wsdl:input name="getSupportedScoringMethods">
677 <soap:body use="literal" />
678 </wsdl:input>
679 <wsdl:output name="getSupportedScoringMethodsResponse">
680 <soap:body use="literal" />
681 </wsdl:output>
682 <wsdl:fault name="PsiscoreException">
683 <soap:fault name="PsiscoreException" use="literal" />
684 </wsdl:fault>
685 </wsdl:operation>
686 <wsdl:operation name="submitJob">
687 <soap:operation soapAction="submitJob" style="document" />
688 <wsdl:input name="submitJob">
689 <soap:body use="literal" />
690 </wsdl:input>
691 <wsdl:output name="submitJobResponse">
692 <soap:body use="literal" />
693 </wsdl:output>
694 <wsdl:fault name="InvalidArgumentException">
695 <soap:fault name="InvalidArgumentException" use="literal" />
696 </wsdl:fault>
697 <wsdl:fault name="PsiscoreException">
698 <soap:fault name="PsiscoreException" use="literal" />
699 </wsdl:fault>
700 </wsdl:operation>
701 <wsdl:operation name="getSupportedDataTypes">
702 <soap:operation soapAction="" style="document" />
703 <wsdl:input name="getSupportedDataTypes">
704 <soap:body use="literal" />
705 </wsdl:input>
706 <wsdl:output name="getSupportedDataTypesResponse">
707 <soap:body use="literal" />
708 </wsdl:output>
709 <wsdl:fault name="PsiscoreException">
710 <soap:fault name="PsiscoreException" use="literal" />
711 </wsdl:fault>
712 </wsdl:operation>
713 <wsdl:operation name="getVersion">
714 <soap:operation soapAction="" style="document" />
715 <wsdl:input name="getVersion">
716 <soap:body use="literal" />
717 </wsdl:input>
718 <wsdl:output name="getVersionResponse">
719 <soap:body use="literal" />
720 </wsdl:output>
721 <wsdl:fault name="PsiscoreException">
722 <soap:fault name="PsiscoreException" use="literal" />
723 </wsdl:fault>
724 </wsdl:operation>
725 <wsdl:operation name="getJobStatus">
726 <soap:operation soapAction="getJobStatus" style="document" />
727 <wsdl:input name="getJobStatus">
728 <soap:body use="literal" />
729 </wsdl:input>
730 <wsdl:output name="getJobStatusResponse">
731 <soap:body use="literal" />
732 </wsdl:output>
733 <wsdl:fault name="InvalidArgumentException">
734 <soap:fault name="InvalidArgumentException" use="literal" />
735 </wsdl:fault>
736 <wsdl:fault name="PsiscoreException">
737 <soap:fault name="PsiscoreException" use="literal" />
738 </wsdl:fault>
739 </wsdl:operation>
740 <wsdl:operation name="getJob">
741 <soap:operation soapAction="getJob" style="document" />
742 <wsdl:input name="getJob">
743 <soap:body use="literal" />
744 </wsdl:input>
745 <wsdl:output name="getJobResponse">
746 <soap:body use="literal" />
747 </wsdl:output>
748 <wsdl:fault name="JobStillRunningException">
749 <soap:fault name="JobStillRunningException" use="literal" />
750 </wsdl:fault>
```

```
751 <wsdl:fault name="InvalidArgumentException">
752 <soap:fault name="InvalidArgumentException" use="literal" />
753 </wsdl:fault>
754 <wsdl:fault name="PsiscoreException">
755 <soap:fault name="PsiscoreException" use="literal" />
756 </wsdl:fault>
757 </wsdl:operation>
758 </wsdl:binding>
759 <wsdl:service name="PsiscoreService">
760 <wsdl:port binding="tns:PsiscoreServiceSoapBinding" name="DefaultPsiscoreServicePort">
761 <soap:address location="http://mint.bio.uniroma2.it/psiscore-ws-0.9.7-SNAPSHOT/webservices/psiscore" />
762 </wsdl:port>
763 </wsdl:service>
764 </wsdl:definitions>
```

Listing 5: PSISCORE Web Service Description Language (WSDL)

2 List of Abbreviations

AIDS	Acquired immune deficiency syndrome
AJAX	Asynchronous JavaScript and XML
AP	Affinity purification
APID	Agile Protein Interaction DataAnalyzer
ATP	Adenosine triphosphate
BIND	Biomolecular Interaction Network Database
BP	Biological process
CC	Cellular component
CHKA	Choline kinase alpha
CSS	Cascading Style Sheet
CV	Controlled vocabulary
DAA	Direct-acting antiviral agent
DAS	Distributed Annotation System
DASMI	Distributed Annotation System for Molecular Interactions
DDI	Domain-domain interaction
DIMA	Domain Interaction Map
DIP	Database of Interacting Proteins
dsRNA	double-stranded RNA
DNA	Deoxyribonucleic acid
DTD	Document Type Definition
DWR	Direct Web Remoting
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENCODE	ENCyclopedia Of DNA Elements
FRET	Fluorescence/Förster resonance energy transfer
GO	Gene Ontology
HCV	Hepatitis C virus
HDF	Host dependency factor
HIMAP	Human Interactome Map
HIV	human immunodeficiency virus
HRF	Host restriction factor
HNRNPK	Heterogeneous nuclear ribonuclear protein K
HPRD	Human Protein Reference Database
HTML	Hypertext Markup Language
HUPO	Human Proteome Organization
HTTP	Hypertext Transfer Protocol
IMEx	International Molecular Exchange Consortium
LUMIER	Luminescence-based mammalian interactome mapping

MAPPIT	Mammalian protein-protein interaction trap
MF	Molecular function
MI	Molecular Interaction
MIQL	Molecular Interaction Query Language
MiMI	Michigan Molecular Interactions
MIMIx	Minimum information required for reporting a molecular interaction experiment
MINT	Molecular Interaction database
MPIDB	Microbial protein interaction database
MPII	Max Planck Institute for Informatics
mRNA	Messenger RNA
MS	Mass spectrometry
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
OBO	Open Biomedical Ontologies
PDB	Protein Data Bank
PI4KIIIa	Phosphatidylinositol 4-kinase III alpha
PINA	Protein Interaction Network Analysis
PPI	Protein-protein interaction
PSI	Proteomics Standards Initiative
PSICQUIC	PSI Common Query Interface
RELAX NX	REgular LAnguage for XML Next Generation
REST	Representational state transfer
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
siRNA	small-interfering RNA
SLM	Short linear motif
SOAP	Simple Object Access Protocol
STRING	Search Tool for the Retrieval of Interacting Proteins
SVR	Sustained virologic response
TAP	Tandem affinity purification
TrEMBL	Translated EMBL
UniHI	Unified Human Interactome
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WAR	Web archive
WSDL	Web Service Description Language
XML	Extensible Markup Language

XSD	XML Schema Definition
Y2H	Yeast-two hybrid

3 List of own publications

Refereed journal publications

1. Alessandra Zanon, Aleksandar Rakovic, **Hagen Blankenburg**, Nadezhda T. Doncheva, Christine Schwienbacher, Alice Serafin, Adrian Alexa, Christian Weichenberger, Mario Albrecht, Christine Klein, Andrew A. Hicks, Peter P. Pramstaller, Francisco S. Domingues and Irene Pichler.
Profiling of parkin-binding partners using tandem affinity purification.
PLoS ONE. 8(11):e78648.1-17, 2013.
doi: 10.1371/journal.pone.0078648.
2. Ágnes Garai, András Zeke, Gergő Gógl, Imre Törö, Fördös Ferenc, **Hagen Blankenburg**, Tünde Bárkai, János Varga, Anita Alexa, Dorothea Emig, Mario Albrecht and Attila Reményi.
Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove.
Science Signaling, 5(245): ra74.1-14, 2012.
doi:10.1126/scisignal.2003004
3. Bruno Aranda*, **Hagen Blankenburg***, Samuel Kerrien, Fiona S L Brinkman, Arnaud Ceol, Emilie Chautard, Jose M Dana, Javier De Las Rivas, Marine Dumousseau, Eugenia Galeota, Anna Gaulton, Johannes Goll, Robert E W Hancock, Ruth Isserlin, Rafael C Jimenez, Jules Kerssemakers, Jyoti Khadake, David J Lynn, Magali Michaut, Gavin O'Kelly, Keiichiro Ono, Sandra Orchard, Carlos Prieto, Sabry Razick, Olga Rigina, Lukasz Salwinski, Milan Simonovic, Sameer Velankar, Andrew Winter, Guanming Wu, Gary D Bader, Gianni Cesareni, Ian M Donaldson, David Eisenberg, Gerard J Kleywegt, John Overington, Sylvie Ricard-Blum, Mike Tyers, Mario Albrecht and Henning Hermjakob. (* equal contributions)
PSICQUIC and PSISCORE: accessing and scoring molecular interactions.
Nature Methods, 8(7):528-529, 2011.
doi:10.1038/nmeth.1637
4. Simon Reiss, Ilka Rebhan, Perdita Backes, Ines Romero-Brey, Holger Erfle, Petr Matula, Lars Kaderali, Marion Poenisch, **Hagen Blankenburg**, Marie-Sophie Hiet, Thomas Longerich, Sarah Diehl, Fidel Ramírez, Tamas Balla, Karl Rohr, Artur Kaul, Sandra Bühler, Rainer Pepperkok, Thomas Lengauer, Mario Albrecht, Roland Eils, Peter Schirmacher, Volker Lohmann and Ralf Bartenschlager.
Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment.
Cell Host & Microbe, 9(3):32-45, 2011.
doi:10.1016/j.chom.2010.12.002

5. **Hagen Blankenburg**, Fidel Ramírez, Joachim Büch and Mario Albrecht.
DASMIweb: online integration, analysis and assessment of distributed protein interaction data.
Nucleic Acids Research, 37(Web Server issue):W122-W128, 2009.
doi:10.1093/nar/gkp438
6. **Hagen Blankenburg**, Robert D. Finn, Andreas Prlić, Andrew M. Jenkinson, Fidel Ramírez, Dorothea Emig, Sven-Eric Schelhorn, Joachim Bøsch, Thomas Lengauer and Mario Albrecht.
DASMI: exchanging, annotating and assessing molecular interaction data.
Bioinformatics, 25(10):1321-1328, 2009.
doi:10.1093/bioinformatics/btp142
7. Andrew M. Jenkinson, Mario Albrecht, Ewan Birney, **Hagen Blankenburg**, Thomas Down, Robert D. Finn, Henning Hermjakob, Tim J.P. Hubbard, Rafael C. Jimenez, Philip Jones, Andreas Kähäri, Eugene Kulesha, José R. Macías, Gabrielle A. Reeves and Andreas Prlić.
Integrating biological data - the Distributed Annotation System.
BMC Bioinformatics, 9(Suppl 8):S3.1-7, 2008.
doi:10.1186/1471-2105-9-S8-S3
8. Michael L. Tress, Pier L. Martelli, Adam Frankish, Gabrielle A. Reeves, Jan J. Wesselink, Corin Yeats, Páll Í. Ólason, Mario Albrecht, Hedi Hegyi, Alejandro Giorgetti, Domenico Raimondo, Julien Lagarde, Roman A. Laskowski, Gonzalo López, Michael I. Sadowski, James D. Watson, Piero Fariselli, Ivan Rossi, Alinda Nagy, Wang Kai, Zenia Størling, Massimiliano Orsini, Yassen Assenov, **Hagen Blankenburg**, Carola Huthmacher, Fidel Ramírez, Andreas Schlicker, France Denoued, Phil Jones, Samuel Kerrien, Sandra Orchard, Stylianos E. Antonarakis, Alexandre Reymond, Ewan Birney, Søren Brunak, Rita Casadio, Roderic Guigo, Jennifer Harrow, Henning Hermjakob, David T. Jones, Thomas Lengauer, Christine A. Orengo, László Patthy, Janet M. Thornton, Anna Tramontano and Alfonso Valencia.
The implications of alternative splicing in the ENCODE protein complement.
Proceedings of the National Academy of Sciences of the United States of America, 104(13):5495-5500, 2007.
doi:10.1073/pnas.0700800104

Book chapters

1. **Hagen Blankenburg** and Mario Albrecht.
PSISCORE (Quality Scoring of Protein Interactions).
In *Encyclopedia of Systems Biology*, Springer New York, ISBN 978-1-4419-9862-0, 1801-1802, 2013.
doi:10.1007/978-1-4419-9863-7_146

-
2. Dorothea Emig, **Hagen Blankenburg**, Fidel Ramírez and Mario Albrecht.
Functional characterization of human genes from exon expression and RNA interference results.
In *Bioinformatics and Drug Discovery (2nd Edition)*, Methods in Molecular Biology, Springer Protocols, Humana Press, New York, NY, USA, ISBN 978-1-61779-964-8, 910:33-53, 2012.
doi:10.1007/978-1-61779-965-5_3

 3. Dmitrij Frishman, Mario Albrecht, **Hagen Blankenburg**, Peer Bork, Eoghan D. Harrington, Henning Hermjakob, Lars Juhl Jensen, David A. Juan, Thomas Lengauer, Philipp Pagel, Vincent Schachter and Alfonso Valencia.
Protein-protein interactions: analysis and prediction.
In *Modern Genome Annotation*, Springer-Verlag, Wien, Austria, ISBN 978-3-211-75122-0, 353-410, 2008.
doi:10.1007/978-3-211-75123-7_17