An Integrated Knowledgebase and Network Analysis Applied on Platelets and Other Cell Types

Integrierte Datenbank und Netzwerkanalysen zur Untersuchung von Blutplättchen und anderen Zelltypen

Doctoral thesis for a doctoral degree at the Faculty of Biology

Julius-Maximilians-University Würzburg

PhD thesis submitted by

Jaya Santosh Chakravarthy Nilla

Würzburg 2012

Submitted on / eingereicht am:

…………………………………………………………………………………………..

Members of PhD student’s committee / Promotionskomitee:

Vorsitzender: Dekan Prof. Dr. Wolfgang Rössler

1. Gutachter: Prof. Dr. Thomas Dandekar
2. Gutachter: Prof. Dr. Christian Wegener

Date of Public Defense / Verteidigung am: ……………………………………………………………..

Date of Receipt of Certificates / Zeugnis erhalten am: ……………………………………………..

Affidavit / Eidesstattliche Erklärung

(According to §4 Abs. 3 Ziff. 3, 5 und 8)

I hereby confirm that my thesis entitled "An Integrated Knowledgebase and Network Analysis Applied on Platelets and Other Cell Types" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Additionally, other than this degree, I have not applied or will attempt to apply for any other degree or qualification in relation to this thesis.

Würzburg

Santosh Nilla

Hiermit erkläre ich an Eides statt, die Dissertation „Integrierte Datenbank und Netzwerkanalysen zur Untersuchung von Blutplättchen und anderen Zelltypen" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Zusätzlich habe oder werde ich nicht versuchen neben diesem Abschluss einen weiteren Abschluss oder Qualifikation mit dieser Doktorarbeit zu erwerben.

Würzburg

Santosh Nilla

For my parents

Contents

Abstract

Systems biology looks for emergent system effects from large scale assemblies of molecules and data, for instance in the human platelets. However, the computational efforts in all steps before such insights are possible can hardly be under estimated. In practice this involves numerous programming tasks, the establishment of new database systems but as well their maintenance, curation and data validation. Furthermore, network insights are only possible if strong algorithms decipher the interactions, decoding the hidden system effects. This thesis and my work are all about these challenges. To answer this requirement, an integrated platelet network, PlateletWeb, was assembled from different sources and further analyzed for signaling in a systems biological manner including multilevel data integration and visualization. PlateletWeb is an integrated network database and was established by combining the data from recent platelet proteome and transcriptome (SAGE) studies. The information on protein-protein interactions and kinase-substrate relationships extracted from bioinformatical databases as well as published literature were added to this resource. Moreover, the mass spectrometry-based platelet phosphoproteome was combined with site-specific phosphorylation/ dephosphorylation information and then enhanced with data from Phosphosite and complemented by bioinformatical sequence analysis for site-specific kinase predictions. The number of catalogued platelet proteins was increased by over 80% as compared to the previous version. The integration of annotations on kinases, protein domains, transmembrane regions, Gene Ontology, disease associations and drug targets provides ample functional tools for platelet signaling analysis. The PlateletWeb resource provides a novel systems biological workbench for the analysis of platelet signaling in the functional context of protein networks. By comprehensive exploration, over 15000 phosphorylation sites were found, out of which 2500 have the corresponding kinase associations. The network motifs were also investigated in this anucleate cell and characterize signaling modules based on integrated information on phosphorylation and protein-protein interactions.

Furthermore, many algorithmic approaches have been introduced, including an exact approach (heinz) based on integer linear programming. At the same time, the concept of semantic similarities between two genes using Gene Ontology (GO) annotations has become an important basis for many analytical approaches in bioinformatics. Assuming that a higher number of semantically similar gene functional annotations reflect biologically more relevant interactions, an edge score was devised for functional network analysis. Bringing these two approaches together, the edge score, based on the GO similarity, and the node score, based on the expression of the proteins in the analyzed cell type (e.g. data from proteomic studies), the functional module as a maximum-scoring sub network in large protein-protein interaction networks was identified. This method was applied to various proteome datasets (different types of blood cells, embryonic stem cells) to identify protein modules that functionally characterize the respective cell type. This scalable method allows a smooth integration of data from various sources and retrieves biologically relevant signaling modules.

Zusammenfassung

Systembiologie sucht nach Systemeffekten in großflächigen Anordnungen von Molekülen und Daten, beispielsweise in menschlichen Blutplättchen. Allerdings kann der Rechenaufwand in den Schritten, die für solche Einsichten nötig sind, kaum unterschätzt werden. In der Praxis umfasst dies zahlreiche Programmieraufgaben, die Einrichtung neuer Datenbanksysteme, sowie deren Wartung, aber auch die Pflege und Validierung der vorgehaltenen Daten. Zudem sind Netzwerkeinsichten nur möglich, wenn effiziente und gute Algorithmen für versteckte Systemeffekte oder auch codierende Wechselwirkungen entschlüsseln. Diese Dissertation und meine Arbeit sind auf diese Herausforderungen konzentriert.

Um diese Anforderung zu erfüllen, wurde ein integriertes Thrombozytennetzwerk, PlateletWeb, aus verschiedenen Quellen zusammengestellt und weiterhin auf Signalverarbeitung und –weitergabe einschließlich mehrstufiger Datenintegration und Visualisierung systembiologisch analysiert. PlateletWeb ist eine integrierte Netzwerkdatenbank, die durch die Kombination von Daten aus den neuesten Thrombozyten Proteom und Transkriptom (SAGE) Studien etabliert wurde. Information über Protein-Protein-Wechselwirkungen und Kinase-Substrat-Paaren wurde aus bioinformatischen Datenbanken hinzugefügt, extrahierte Daten aus der veröffentlichten Literatur ergänzten dies weiter. Darüber hinaus wurde das Blutplättchen-Phosphoproteom aufgrund von Daten aus der Massenspektroskopie mit ortsspezifischen Phosphorylierungs- / Dephosphorylierungsdaten kombiniert. Ergänzt wurde dies um Daten aus der Datenbank Phosphosite und durch bioinformatische Sequenzanalyse unter Nutzung ortsspezifischer Kinasevorhersagen. Die Zahl der katalogisierten Thrombozytenproteine wurde im Vergleich mit der Vorversion von 2008 um mehr als 80% erhöht (beinahe Verdoppelung der Daten, insbesondere aber neue, zusätzliche Datenkategorien, z.B. über Pharmaka, Phosphorylierung, Gen-Ontologie, daneben auch weitere Validierung und Pflege der vorhandenen Daten). Die neue Integration von Annotationen für Kinasen, Proteindomänen, Transmembranregionen, Gene Ontology, Krankheitsbezüge und Azneimittelziele bietet

neue, mächtige Werkzeuge für die funktionelle und systembiologische Analyse von Thrombozytensignalwegen. Die PlateletWeb Datenbank liefert eine neuartige systembiologische Werkbank zur Analyse von medizinisch relevanten Blutplättchensignalen (z.B. Plättchenaktivierung bei Thrombose, Hämostase etc.) im funktionellen Zusammenhang von Proteinnetzwerken. Durch umfassende Untersuchungen wurden über 15000 Phosphorylierungsstellen identifiziert, von denen 2500 einer Kinase zugeordnet werden konnten. Netzwerkmotive wurden auch in diesen Zellen ohne Zellkern untersucht und neue und interessante Signalmodule charakterisiert. Dies war nur durch die integrierte Information über Phosphorylierung und Protein-Protein-Wechselwirkungen möglich.

Darüber hinaus wurden zahlreiche algorithmische Ansätze verwand, darunter ein exakter Ansatz zur Bayesschen Analyse von Interaktionsnetzwerken (Heinz) basierend auf linearer Integer-Programmierung. Gleichzeitig hat sich unser Konzept der semantischen Ähnlichkeiten zwischen zwei Genen basiert auf Gene Ontology (GO) Annotationen etabliert und ist eine wichtige Grundlage für viele analytische Ansätze in der Bioinformatik geworden. Unter der Annahme, dass eine höhere Anzahl von semantisch ähnlichen funktionellen Genannotationen biologisch relevantere Interaktionen reflektieren, wurde eine Bewertung der Kanten für funktionelle Netzwerkanalyse entwickelt.

Die Kombination beider Ansäte, die Kantenbewertung, basierend auf der GO-Ähnlichkeit und die Netzknotenbewertung bezogen auf die Expression der Proteine ermöglichte in den analysierten Zelltypen (unter Nutzung von Daten z.B. aus Proteomstudien) die Identifizierung funktioneller Module als maximal bewertete Subnetzwerke in großen Proteinnetzwerken. Dieses Verfahren wurde an verschiedenen Proteomdatensätzen getestet (verschiedene Arten von Blutzellen, embryonale Stammzellen), um Proteinmodule zu identifizieren, die funktionell den jeweiligen Zelltyp charakterisieren. Weitere Ansätze der Methode erfassen die Analyse von quantitativen Phosphoproteom-Daten zur Identifizierung des Signalflusses in einem Kinase-Substrat Netzwerk. Diese skalierbaren Ansätze ermöglichen eine reibungslose Integration von Daten aus verschiedenen Quellen und liefern biologisch relevante Signalmodule.

1 Introduction

Understanding the systems biological background of cellular processes is crucial for instance, identifying potential new biomarkers in human diseases. Genomic and proteomic data were generated in recent years by new tools and technologies. The focus is to integrate the information from multiple sources and manage the data with high reliability. The integrated data should have a backward tracking which would help identify every protein's identification evidence and its functional information. This ever increasing number of proteomic and phosphoproteomic data generated by mass spectrometry calls for new strategies in data organization, representation and analysis.

A cell type of particular interest is the blood platelet for its role in thrombosis and hemostasis and cardiovascular disease, which is still the major cause of all deaths. Platelets (also called as thrombocytes) are small, irregularly shaped clear cell fragments which do not have a nucleus. The average life span is around 5 to 9 days. They circulate in the blood freely, if the number of platelets is too low, then this might result in excessive bleeding. However, if the platelet count is too high, blood clots can form and cause serious disease such as stroke or active heart failure. Platelets thus play a key role in hemostasis and represent a central target for research in many pathophysiological processes, including cardiovascular diseases, inflammatory processes and metastasis (*Varga-Szabo, Pleines et al. 2008*).

In this thesis project, an integrated platelet interactome was assembled from multiple sources based on the manually curated protein data and the interactions between each of them. This platelet interactome could contain keys to solving the exact mechanisms of platelet activation and aggregation. Understanding the underlying mechanisms of platelet activation could hold the answer to new treatments and potential drug targets. Also, platelets are important anucleate cells where only posttranscriptional regulation is possible. A tight regulation of protein interactions is needed to ensure proper signaling. A thorough analysis of the platelet proteome and interactome provides a better understanding of signaling principles in anucleate cells.

A huge amount of mass spectrometry data on platelet proteins has been accumulated in recent years. However, a platform to combine both the phosphorylation and interaction information of these proteins is missing. The PlateletWeb is developed in order to overcome this factor and thus is the resource where the complete information about the platelet proteins was easily accessible. Furthermore, the integration of various data sources proves useful for a complex systems biological analysis of a given cell system. The proteome and transcriptome information on platelet proteins was combined along with phosphorylation and interaction data, thereby assembling a human platelet interactome and phosphoproteome containing platelet kinases and phosphorylations from literature and experimental sources. Kinase predictions for experimentally validated phosphosites were also included in the network (*Linding, Jensen et al. 2007; Miller, Jensen et al. 2008*). The newly created interactome was complemented by drug target information, disease associations, gene ontology annotation and KEGG pathway data allowing a more thorough analysis of platelet signaling. Finally, an integrated network of human platelets was established covering the proteome, phosphoproteome and transcriptome based on data from mass-spectrometry studies, public databases and literature research as described in Materials and Methods. Careful annotation resulted in a set of 5025 platelet proteins with evidence on either proteome or transcriptome level.

Additional to this, in order to make the PlateletWeb a complete resource for platelet information, the data from DrugBank and KEGG were included. The drug database associated with the protein along with the genetic diseases provides a simple understanding of various levels of the protein information otherwise a tedious task to investigate. The KEGG pathways and the KEGG orthology were also provided focusing basically on the platelet proteins. Furthermore the transmembrane domain predictions, the protein characteristics, the Gene ontology information etc. on each of the platelet protein would make the PlateletWeb resource, a complete systems biological workbench providing an unmatched information about the platelets at one single place.

Based on this information a study on other cell types was initiated to investigate the role of similar proteins and similar pathways in multiple species. An interactome of mouse proteins was established and this was then mapped to the human proteins in order to find the orthologs and homologs of the proteins. Similar to the PlateletWeb database (*Boyanova, Nilla et al. 2012*), complete protein information was also added to the mouse database. A first draft of information was initiated, which forms the basis of a mouse platelet database to be published later. Furthermore, it can be updated to many other species and look for the protein of interest from multiple aspects.

Additionally, the complete cell type analysis was made possible using the heinz algorithm included with the GOSim analysis. I could make the best use of the well curated Gene Ontology information and impute it into the heinz algorithm to quantify the proteins and their interactions. This helped to retrieve networks which are of high biological relevance. The complete analysis requires a higher computational details and insights with comparatively a higher challenge on the algorithm preparation. This GOSim algorithm design is mainly focus on the key aspect of it working with any sample proteins that are identified in any proteomic study.

Furthermore, the motif analysis was performed to see the network patterns with in the constructed network of platelet interactome and identify the enrichment of a specific sub graphs with in the network. This was helped to identify the key switches in the network module and there by interpreting the information from the statistical overview.

The available gene ontology information provided for all the proteins would help in understanding the protein in terms of their biological process, their molecular functionality and the place in the cell where it is located in the human readable format. This information is taken from the GO consortium and assembled to the platelet proteins. The gene enrichment analysis when performed, this human readable gene ontology would provide first insights on what the protein cluster does.

The interaction between two proteins is specified in multiple ways; however, the quantification of the relation between two proteins would be a further step in achieving the similar clusters of proteins which perform a similar task. This quantification is performed by taking the gene ontology information between two proteins which are interacting between each other and then checking for the semantic similarity between them and there after quantifying it. This would help to extract and understand the modules from the complete network of proteins which are specific to a certain criteria. A complete testing was performed by taking already existing modules (to cross check the validity of the extracted modules) and also by trying to figure out the new modules in different proteins identified in multiple samples.

The quantitative and qualitative analysis which was carried out further helped to get the insights of the proteins which are otherwise might be overlooked.

Thus PlateletWeb is the key resource for understanding, analyzing and predominantly a highly capable source of information about the platelet proteins giving the complete insight of each protein individually as well as the role of protein within a specific network module and the role of the cluster of proteins as a whole. Thus the PlateletWeb could be rightly termed as the “Systems biological workbench”.

2 Materials and Methods

2.1 Proteomic databases

The information about the platelet proteins, their interactions, phosphorylations, kinase information, drugs, genetic diseases information along with the gene ontology information is taken from multiple databases. The main source of information was the proteomic databases; the main focus is on the manually curated data, which would provide with the high end reliable information. The proteomic databases used to compile the data are given below.

2.1.1 HPRD

The Human Protein Reference Database is a database of manually curated proteomic data focusing mainly on human proteins. This database was initially assembled in 2003 with notable updates there on (*Peri, Navarro et al. 2003*). This database contains relatively a small set of 2,750 proteins and 25,050 literature references. The current version (HPRD release 9) (*Keshava Prasad, Goel et al. 2009*) contains over 30,000 proteins, with more than 40,000 protein-protein interactions and over 90,000 post translational modifications. These statistics are obtained after careful annotations performed by biologists based on literature analysis of more than 450,000 PubMed links. Data is manually curated and no text mining algorithms are used. This sophisticated process ensures a very high quality of protein data coming from published articles.

2.1.2 Other databases

Information on human proteins and their interactions were downloaded from Entrez gene Database (NCBI)(*Maglott, Ostell et al. 2007*). NCBI provides a unique identifier called Entrez gene identifier (also called GeneID) where every gene is assigned a unique number. This ensures a proper cross link between multiple databases there by not repeating the same gene twice even though taken from multiple sources. All proteins in the complete *PlateletWeb* database are identified uniquely on the basis of gene identifiers.

2.1.3 Individual Platelet Studies

The Platelet proteome catalogue was first defined by Dittrich et al (*Dittrich, Birschmann et al. 2008*) which includes an assembly of a comprehensive proteome and transcriptome database of human platelets. A model of the platelet interactome was later created with the platelet proteins. Based on this catalog of the platelet proteome, further data from various mass spectrometry studies published over the last 10 years have been assembled. Most of these studies are mainly focusing on the unfractionated platelets, specific platelet subcompartments including the plasma membrane, secretome and microparticles. Additionally, a previously performed Serial Analysis of Gene Expression (SAGE) study of human platelets was included. Also literature curated information was extracted from the NCBI GeneRifs (*Maglott, Ostell et al. 2007*) and were filtered for new platelet proteins. This comprehensive set of platelet proteins from multiple sources resulted in 5025platelet proteins.

2.1.4 Kinases and Phosphatases

A comprehensive list of human kinases were extracted from Manning et al (*Manning, Whyte et al. 2002*) and used for reference and validation of the HPRD phosphorylation data. A kinase is a type of enzyme that transfers phosphate groups from high-energy donor molecules, such as ATP to specific substrates. This process is called phosphorylation and is important to identify the dynamics in the protein network regulation. All kinases were mapped to the human kinome tree created by Linding et al (*Miller, Jensen et al. 2008*) and visualized using the interactive online tool Tree Of Life (*Letunic and Bork 2007*). The catalogue of human phosphatases was acquired from the Human Protein Phosphatases PCR Array (Quiagen; 82 phosphatases) and the assembly of protein tyrosine phosphatases in the human genome (*Alonso, Sasin et al. 2004*) (103 phosphatases). The rest of the phosphatases were added by manual search in the *PlateletWeb* for proteins with the term “protein phosphatases” in their description. The total number of human protein phosphatases adds up to 191 (phosphatases associated with a substrate: 39, platelet phosphatases: 73, platelet phosphatases with a substrate: 24).

2.2 Protein-protein interactions (PPI)

A considerable number of databases came into existence in the recent past focusing mainly on the protein-protein interactions, many of which are emphasizing on more than one species (*Wang, Nath et al. 1999; Goddard, Ladds et al. 2006; Stark, Breitkreutz et al. 2006; Stark, Breitkreutz et al. 2011*) give a few citations here: BioGrid, BIND and all the rest). The core idea of *PlateletWeb* is to create a network of human platelet proteins for detailed analysis on a single and multiple protein level. In order to accomplish this, the information from multiple databases was taken into consideration. The manually curated human protein-protein interactions were extracted from the Human Proteome Reference Database (HPRD) (version 9.0, 04/2010) (*Keshava Prasad, Goel et al. 2009*), which contains 39,194 simple binary PPIs and 93,710 post-translational modifications. The post-translational modifications were also considered for the interactions list, as the modification itself is possible only if there exist an interaction between the proteins. Information about the type of experiment, in which the interaction was found, is also presented with each of the interaction. Two proteins are considered to interact *in vivo* when the interaction was detected in a mammalian cell at the time of experiment. If it could not be concluded in the context of mammalian cells, it was considered *in vitro*. The third type of experiment is the Yeast 2-hybrid which describes interactions only detected in a yeast cell after performing this type of experiment. These can be technically termed as *in vivo*, but were annotated specifically as Y2H in order to identify the false positives interactions, if any. HPRD also includes data from Biological General Repository for Interaction Datasets (BioGrid) (*Stark, Breitkreutz et al. 2011*) and Biomolecular Interaction Network Database (BIND) (*Alfarano, Andrade et al. 2005*). Each interaction is also marked with its interaction sources thus helping to trace back to its source study.

The data was downloaded from the HPRD server in the FLAT file format. This file was then parsed using the Perl program and created as a table in the MySQL database. With the help of this structured query language, it was possible to remove the duplicate interactions and maintain the data in a definite order. The HPRD provides its proteins with a unique identifier called HPRD Id. The Entrez gene identifiers were them mapped to maintain consistency and ease of use.

Another database that was taken into consideration with a particular interest is Phosphosite (accessed 01/2011) (*Hornbeck, Chabra et al. 2004*). Phosphosite is a systems biological resource which provides information about the protein post translational modifications. The database is manually curated and provided over 3000 protein translational modifications many of which are additional to the ones listed in the HPRD database. Additionally, an additional set of post translational modifications (both phosphorylations and dephosphorylations) on Serine (S) Threonine (T) and Tyrosine (Y) phosphorylations were included as a table in MySQL.

A set of phosphorylations were also complemented by kinase predictions for platelet-specific phosphoproteome data (*Zahedi, Lewandrowski et al. 2008*) using the NetworKIN algorithm (*Linding, Jensen et al. 2007; Miller, Jensen et al. 2008*). NetworKIN is a method for predicting *in vivo* kinase-substrate relationships. This algorithm combines two different approaches for phosphorylation prediction - consensus sequence motif search and protein association networks (a network context of kinase and phosphoproteins which makes up to 60-80% of computational capability to assign in vivo substrate specificity) in order to generate a full, realistic and statistically more probable prediction of the involved kinase. An online website was provided to predict the kinase substrate relationships (http://networkin.info/version_2_0/search.php). Results from experimental mass spectrometry analysis with determined phosphorylation sites were inserted into the website's search functionality, which predicts the kinase responsible for a specific phosphorylation. Two different scores (a motif score and a context score) are calculated for each algorithm and presented in the final results. The predicted kinase substrate relationships were also taken into consideration as additional information into the PPI network.

In total, the complete human PPI network contains 54,218 simple interactions, 4,406 phosphorylation events and 135 dephosphorylation events between 10,916 human proteins.

2.3 Network motif analysis

Networks are complex structures which are defined by nodes and interconnected edges. These networks carry a pattern which when recognized could provide an in-depth of network regulation and signaling. To unravel these networks, Milo et al (*Milo, Shen-Orr et al. 2002*) defined the term "Network Motifs". These network motifs are the patterns occurring in the complex network with numbers higher than those in the randomized networks.

The Mfinder Version 1.2 software was used in order to enumerate the specific network motifs in the interactome and phosphoproteome network. The enrichment of a specific motif within the network is thereby calculated and helps to investigate various platelet signaling pathways.

The software Mfinder Version 1.2(*Milo, Shen-Orr et al. 2002; Kashtan, Itzkovitz et al. 2004*) was used for detection of specific predefined(*Zaidel-Bar, Itzkovitz et al. 2007*) network motifs in the combined platelet interactome and phosphorylation network and analysis of overrepresentation as previously described by Milo et al(*Milo, Shen-Orr et al. 2002*). The algorithm identifies all n-node subgraphs in the input network. A set of randomized networks (100 randomizations were used in this analysis) was created by rewiring the edges but maintaining the same incoming and outgoing degree at each node. All randomized networks are examined for network patterns in the same way as the real network (=biological network) and the number of occurrences is calculated in each case. Z-scores are obtained by subtracting the mean motif count in the randomized network from the observed count in the real network and subsequently dividing the result by the standard deviation of the randomized motif count (The Z-scores are defined with the following formula: (Nreal – Nrand)/SD (*Milo, Shen-Orr et al. 2002*)). These final scores indicate the significance of enrichment for each motif in the real network. Subgraphs with high absolute Z-scores tend to be significantly enriched and are regarded as network motifs(*Milo, Shen-Orr et al. 2002*).

2.4 Drugs and diseases

Drug data were downloaded from DrugBank Version 3.0 (*Knox, Law et al. 2011*), which includes detailed information on drugs as well as on drug targets. The physical drug-target interactions are identified to get the overview on the possible indirect functional effects. The drugs are divided into multiple broad categories, mainly in approved and experimental groups. The database contains 4311 human drugs, which have a human drug target in the *PlateletWeb* knowledge base (approved, 1195; experimental, 3015) and act on 2106 distinct human proteins. There are 950 platelet proteins among these drug targets. Genetic disease information was extracted from HPRD and is available for 701 platelet proteins.

2.5 KEGG Pathways

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a Pathway database. It is a collection of databases dealing with genomes, enzymatic pathways, and biological chemicals. The KEGG pathways were downloaded from the KEGG database (Release 57.0, January 1, 2011). However, from the middle of 2011, KEGG has switched to a subscription model and access via FTP is no longer free. Each pathway was termed individually to identify the platelet proteins. The platelet proteins identified in the pathways were visualized in the *PlateletWeb* knowledge base using the Advanced Pathway Painter v2.26. Enrichment analysis of pathways was performed using Fisher's exact test comparing the number of platelet proteins in the pathway against the number of all platelet proteins annotated in KEGG pathways.

2.6 Gene Ontology

In order to understand their biological role, genes are annotated by scientists using human language. This would help to understand the purpose of genes in a natural language format. However, when taking about the large-scale proteomics, there is a high probability that the genes are redundantly annotated with fair amount of ambiguity. In order to overcome this issue, the Gene Ontology consortium (*Ashburner, Ball et al. 2000; 2008*) has been developed which strives for a well-defined and structured functional assignment of genes. The GO terminology is arranged in a hierarchical way, which is also termed directed acyclic graph (DAG). As one goes down the hierarchy, terms become more specific. A parent GO term can be then sub divided into one or many different child terms, which describe a much more specific biological role than the general term.

The GO is divided into three branches, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). The BP tree defines the overall process in which the gene product is involved, MF specifies the biochemical function and CC denotes the compartment or subcellular stricter where the gene product is located. As an example, the term "apoptosis" is identified in Biological Process, "kinase activity", a molecular function and the term "nucleus" belongs to the ontology cellular component.

Each biological function in the gene ontology is associated with a GO term, which is a unique identifier. Also, each of the go terms can be associated with one or more proteins and each of the proteins can be assigned to more than one go terms. The GO can be accessed as a downloadable file format or in the form of the MySQL table structure, where the required data can be easily queried. Alongside, there are a large number of tools available along with the GO which would help in elucidating the functional role of proteins.

The vocabulary in GO has been constantly growing, with over 35,774 terms defined as on February 2012. Among them, the biological process has 21,964 terms defined; the molecular function has 9,238 and 2,960 in cellular component.

The whole database was downloaded and filtered for human proteins from GO and this was used for the functional profiling of the proteomic data. There are 4,728 platelet proteins annotated with a GO function, which accounts for a coverage of 94%.

2.7 Semantic Similarity

The concept of semantic similarity is mainly based on the very key notation, when there exist 2 nodes A and B and the main aim is to determine how similar these nodes are to each other. This can be quantified ranging between the values of 1 and 0. A value of 1 represents high similarity between the nodes and a value of 0 represents either no similarity or extremely low similarity.

This concept was applied to proteins and their biological ontologies. The valuation for the semantic similarity between two genes is based on the Gene Ontology terms that are common between these genes. The scores for the semantic similarity were initially calculated based on the method defined by Schlicker et al (*Schlicker, Domingues et al. 2006*) where the probability of the most informative common ancestor (MICA) is defined. The function of semantic similarity between "two terms" is defined as:

$\mathbf{(Sim_{Rel}(C_1,C_2) = Sim_{Lin}(C_1,C_2) \times (1\text{-}C_{MICA}))}$ (*Schlicker, Domingues et al. 2006*),

where C1 and C2 represent two GO terms. This algorithm was further extended by Frohlich et al (*Frohlich, Speer et al. 2007*) to calculate functional similarity between "two genes" (GOSim). The *getGeneSim* function in the GOSim package along with the *funSimAvg* as similarity measure (*Schlicker, Domingues et al. 2006*) determines the average of best matching GO term similarity for both genes. The semantic measurement was calculated for all the three ontologies (BP, MF and CC) on all interactions listed in the interactome. These results were then combined together for each interaction into one composite score using the BioNet package in R (*Beisser, Klau et al. 2010*).

2.8 Heinz algorithm extended with semantic similarity

Recent experimental advancements are allowing the detection of the proteome which is either expressed in the cell or in a particular compartment of the cell. Once the proteins are identified, it's a challenge to interpret them in the biological context. The analysis of a network obtained by combining data from expression profiling study of lymphoma patients with the comprehensive interactome data from HPRD was performed previously by Dittrich et al (*Dittrich, Klau et al. 2008*). In this prospect, the p-values are derived from the analysis of differential expression between two tumor subtypes as well from the analysis of survival data by cox regression for each node in the interaction network. The main idea is to identify functional modules in the PPI network, sharing common cellular functions. In order to achieve this, a maximally scoring network is devised along with the scoring of the nodes in the network to be identified.

This identification of the functional modules in the network was extended by weighing the edges depending upon the semantic similarity of their nodes. If two genes are semantically similar to each other they would get a higher value close to or equal to 1 and if they are not semantically similar, they would get a value close to 0 or equal to 0. These scores were calculated using the "GOSim package" for all the three ontologies – BP, MF and CC. The scores were then converted into empirical p-values and using the aggregation statistics previously defined by Dittrich et al (*Dittrich, Klau et al. 2008*), which was based on distribution of the order statistics, a single p-value was obtained. Based on these aggregated p-values, a scoring function was defined which represents the p-values as a mixture of a noise and a signal component. The scores obtained from this scoring function were termed "Edge Scores". Once the edge scores were calculated, the nodes of interest (proteins identified in the sample) were scored by calculating the negative average of the scores of their connected edges. The rest of the nodes were given the average of all edge scores.

The BioNet package (*Beisser, Klau et al. 2010*) provides an extensive framework for integrated network analysis using R and BioConductor. The methods for finding the optimal and suboptimal solutions were extensively used for finding the functional networks.

In order to identify the functional modules, an extensive usage of heavy induced graphs (heinz) was made. The algorithm to identify the optimal scoring subnetwork is based on the software dhea (district heating) from Ljubic et al. (*Ljubic, Weiskircher et al. 2006*). The program is developed in C++ to generate the optimal and suboptimal solutions. This is controlled over a Python script. The dhea code uses the commercial CPLEX license to calculate the suboptimal solutions. The dhea code and heinz python script are publicly available for academic and research purposes within the heinz (heaviest induced subgraph) package of the open source library, LiSA (http://www.planet-lisa.net). If the CPLEX license is not available, the maximum-scoring subnetwork can still be calculated with the included heuristic in the BioNet package (*runFastHeinz*).

2.9 Phosphoproteomics (Quantitative analysis)

Phosphoproteomics is one of the rapidly evolving fields in mass-spectrometry analysis. In contrast to previous methods which mainly rely on whole-protein data, this approach focuses on the analysis of cell-wide phosphorylation patterns. This integrated analysis combines protein-protein interaction (PPI) networks along with phosphoproteomic data to functionally describe signaling pathways and the change of information flow during various states of stimulation. Quantitative phosphoproteome data for node scoring in networks derived from PPI data along with the kinase-substrate relationships is used to understand this in the systems biological manner. Additional kinase information is integrated into the network as it is essential for understanding the regulatory mechanisms of protein signaling.

The maximum-scoring subnetwork using an exact approach (heinz) is used to identify differentially phosphorylated signaling modules in cellular networks. This exact approach for searching biologically relevant functional modules has already been introduced in microarray analysis and is now extended to site-specific phosphoproteomics data. The algorithm in combination with human kinase-substrate relationships from *in vivo* and *in vitro* experiments on a dataset of differentially phosphorylated human embryonic stem cells (hESCs) (*Rigbolt, Prokhorova et al. 2011*) was used. With this approach, the analysis of networks under various conditions (including time-series experiments) is carried out, thereby characterizing system states in a network context.

A total of 6,521 proteins were identified in the original dataset (*Rigbolt, Prokhorova et al. 2011*) using Stable isotope labeling by amino acids in cell culture (SILAC) method, of which 5765 proteins were mapped in the *PlateletWeb* database along with 205 kinases. The site-specific phosphorylation changes were measures after application of non-controlled medium (NCM) 30 minutes, 1 hour, 6 hours and 24 hours after stimulation. 12,070 distinct peptides with 141 kinases acting on these phosphosites were identified. Of these, 539 phosphosites and 281 proteins were associated with at least 1 kinase. The SILAC ratios were calculated in order to get the differential phosphorylation between the treated and the controlled cells.

The SILAC ratios were then transformed into the site specific node scores and used the algorithm to obtain time-specific response modules of phosphorylation signaling during hESCs differentiation.

2.10 Visualization and other tools

TMHMM: Transmembrane domains have been predicted using the TMHMM Server, Version 2.0 (*Krogh, Larsson et al. 2001*) yielding a total of 5,107 transmembrane proteins, of which 1,158 are platelet proteins

Cytoscape: Cytoscape is an open source platform for complex network analysis and visualization. It is a java based tool and can be run as standalone software on the local computer. Throughout the study, the visualization of the subnetworks is performed by Cytoscape version 2.6.3, unless specified.

Bingo Plugin: Bingo (Biological Network Gene Ontology Tool) is a plugin for Cytoscape. It helps to determine which Gene Ontology categories are statistically over or under represented in a set of genes. Gene enrichment analysis was performed by the BINGO plug-in v2.44 of the network analysis software Cytoscape 2.8. For the full GO annotation comparison, all platelet proteins with a GO functional annotation in the network were considered (Biological Process (BP): 3,263; Molecular Function (MF): 3,412; Cellular Component (CC): 3,394 of total 5,025 platelet proteins). Statistically significant categories (P<0.0001) were selected according to their corrected p-values, using a hypergeometric test. Visualization of the BP, MF and CC results was performed for selected GO terms with less than 600 proteins. The top 25 GO terms were then used for visualization and colored according to the common parent with high information content in the hierarchical GO tree.

R: R is a language for statistical computing and graphics. It is free software under the terms of GNU in the source code form. R was used vigorously throughout the project for multiple analyses after installing the required packages.

Igraph: "igraph" is a free software package for creating and manipulating undirected and directed graphs. This package was used for the online visualization (on the PlateletWeb website) of networks on a chosen subset of proteins. This package can be used in multiple forms, for this work, R package was installed and used.

MySQL: MySQL is a relational database management system and it's a structured query language. The data taken from multiple sources were saved in the form of tables in the database and uniqueness was maintained for the gene identifiers.

PHP/CSS YAML framework: The *PlateletWeb* website was designed using the PHP (Hypertext Preprocessor) with an extensive usage of CSS (Cascaded Style Sheets) and its framework, YAML (Yet Another Multicolumn Layout). MySQL was used as the backend for the website.

Perl: Parsing of sequences and flat files were performed using Perl (Practical Extraction and Reporting Language).

Microsoft Office; Notepad++: Typesetting of manuscripts and periodical updates were saved in Microsoft Word and Notepad++. Additionally, the Microsoft Excel was used for a first overview of the processed data in the graphical formats.

XAMPP: XAMPP (Cross platform, Apache, MySQL, PHP and Perl) is a free and open source cross platform webserver package. XAMPP is extensively used for the development version of the *PlateletWeb* knowledgebase and the *PlateletWeb* resource.

3 Results

3.1 PlateletWeb

Understanding the cellular mechanisms of platelet activation and their pharmacological modulation is of major interest in cardiovascular platelet research. A first step towards obtaining deeper insight into platelet signaling networks is to look at key molecular building blocks. Therefore, a multi-functional platelet network database (*PlateletWeb*) was established by collecting proteome data from large-scale proteome studies, well-curated protein databases, published literature as well as from transcriptome data (*Dittrich, Birschmann et al. 2005; Dittrich, Birschmann et al. 2008*). *PlateletWeb* provides a novel systems biology workbench for the analysis of platelet signaling in the functional context of integrated networks. An integrated network database was established combining data from recent platelet proteome and transcriptome (SAGE) studies with information on protein-protein interactions and kinase substrate relationships extracted from bioinformatical databases as well as published literature. Moreover, mass spectrometry-based platelet phosphoproteome was combined with site-specific phosphorylation / dephosphorylation information from the Human Protein Database (HPRD) and Phosphosite and complemented by bioinformatical sequence analysis for site-specific kinase predictions. The number of catalogued platelet proteins was increased by over 80% as compared to the previous version (*Dittrich, Birschmann et al. 2008*). Integration of comprehensive annotations on kinases, protein domains, transmembrane regions, Gene Ontology, disease associations as well as drug targets provides ample functional tools for platelet signaling analysis. The resource is made available as PlateletWeb knowledgebase and this can be reached from the website *http://plateletweb.bioapps.biozentrum.uni-wuerzburg.de/plateletweb.php*

3.1.1 Data sources

The PlateletWeb database is a highly scalable, reliable and up to date database, which has the tendency to expand easily with more information that comes in. My part in the study was to make sure and implement this database with the huge consideration on various factors, for instance scalability, non-redundancy, consistency etc. In order to achieve this, the structure of the database should be consistent throughout and the addition of information from any source would not affect the already existing information. In order to achieve this, a careful analysis on the proteins from multiple sources was made and similar proteins from multiple databases are mapped to gain a highly reliable and sophisticated database which contains unique protein coming from multiple sources. The assembly of this information requires a high detailed overview on the source data that is being used. A careful annotation of each of these proteins is also required in some special cases where the protein in the source is wrongly annotated.

Platelet proteome data was assembled from recent mass spectrometry studies of unfractionated platelets and specific platelet subcompartments. Platelet proteins were gathered from various studies including the latest proteomic whole cell and secretome analysis to ensure a complete set of new proteins. Additionally the data was taken from multiple data sources which have the human protein specific information. The information was combined and emphasis was given on the platelet proteins.

Platelet proteome data have been collected manually from all available literature and proteomic databases and combined with transcriptome data (SAGE). It resulted in a set of 5025 platelet proteins, including 22 major platelet datasets, creating a comprehensive and reliable backbone for platelet-specific information.

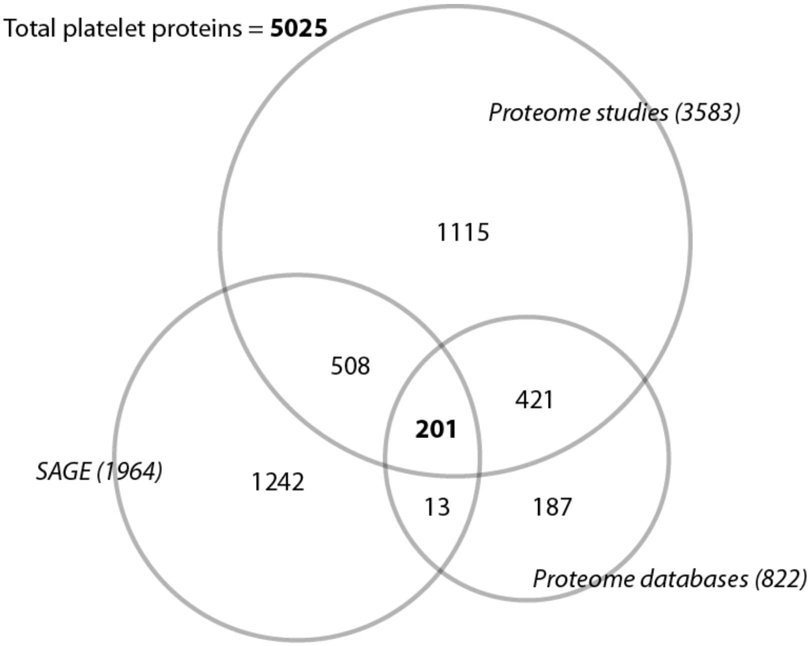


Figure 1 - Proteins extracted from literature and databases

Different sources are compared in this Venn diagrams, e.g. 201 proteins are found by all three methods. The intersections and specific numbers show that the different sources are rather complementary then redundant.

Proteins were categorized according to the cellular subcompartments from which they were isolated, which revealed that membranes and secretome are the most abundant fractions. Proteins were analyzed individually depending on the source of platelet information (proteome and transcriptome). The proteome group was in turn subdivided into proteome studies (large-scale proteomic analysis) and proteome databases (bioinformatic databases). The majority of proteins (3783 proteins, 75%) have been described on the proteome level, 14% (722 proteins) have additional evidence on the transcriptome level, and 25% (1242) of all platelet proteins have been detected exclusively on the mRNAlevel.

A detailed analysis of the platelet protein sources was performed to identify the studies and fractions, from which the platelet proteins were extracted. The largest part of platelet proteins are coming from the meta database of mass spectrometry data GPMBD (citation). The transcriptome represents the second most abundant source of platelet information.

Interestingly, the membrane platelet proteome study by Lewandrowski et al contains the highest number of extracted platelet proteins in a single study. A complete list of platelet proteins along with their study and fraction is provided in the table below.

Table 1 - Number of platelet proteins extracted from all studies:

Study/Database	***Number of proteins***	***Fraction***
GPMBD 2011	2064	whole platelet
SAGE(Dittrich, Birschmann et al. 2006)	1964	whole platelet
Lewandrowski 2009(Lewandrowski, Wortelkamp et al. 2009)	1269	membrane
Piersma 2009(Piersma, Broxterman et al. 2009)	707	secretome
Haudek 2009(Haudek, Slany et al. 2009)	671	whole platelet
Martens 2005(Martens, Van Damme et al. 2005)	608	whole platelet
Garcia 2005(Garcia, Smalley et al. 2005)	554	microparticles
Uniprot 2009(2009)	538	undefined
Wong 2009(Wong, McRedmond et al. 2009)	358	whole platelet
Garcia 2004(García, Prabhakar et al. 2004)	305	whole platelet
Moebius 2005(Moebius, Zahedi et al. 2005)	281	membrane
Thon 2008(Thon, Schubert et al. 2008)	279	whole platelet
Zahedi 2008(Zahedi, Lewandrowski et al. 2008)	277	phosphoproteo

Study/Database	***Number of proteins***	***Fraction***
		me
HPRD(Keshava Prasad, Goel et al. 2009)	230	undefined
Maynard 2007(Maynard, Heijnen et al. 2007)	206	alpha granules
Coppinger 2004(Coppinger, Cagney et al. 2004)	182	secretome
Guerrier 2007(Guerrier, Claverol et al. 2007)	176	whole platelet
Coppinger 2007(Coppinger, Fitzgerald et al. 2007)	132	secretome
GeneRIF	120	undefined
Springer 2009(Springer, Miller et al. 2009)	119	whole platelet
O'Neill 2002(O'Neill, Brock et al. 2002)	116	whole platelet
Marcus 2000(Marcus, Immler et al. 2000)	109	whole platelet
Garcia 2006(Garcia, Senis et al. 2006)	77	whole platelet
Yu 2010(Yu, Leng et al. 2010)	77	whole platelet
Thiele 2007(Thiele, Steil et al. 2007)	29	whole platelet
Glenister 2008(Glenister, Payne et al. 2008)	16	whole platelet

All platelet information sources are presented with the source of data, number of platelet proteins and the fraction from which the proteins were extracted. Each platelet protein was identified in at least one or several studies, this information was provided on the PlateletWeb website under “Platelet Evidence”.

Analysis indicates that whole platelet investigation is predominantly used in platelet proteomics. This resource helps in understand information about the complete platelet easily along with its interactions and phosphorylations with other platelet proteins as well as the human proteins. Single compartment analysis such as membrane or microparticles is difficult to achieve due to technical limitations in extraction of specific platelets sub compartments.

3.1.2 PlateletWeb Statistics

The analysis on any of the databases reveals the amount of information it contains and this can be easily achieved by using the statistics. Additionally, the numbers would help to maintain the integrity in the complete network and it easily shows the flaws if for instance, some proteins or some interactions are falsely identified (false positives or false negatives). However, to gain the statistical analysis, it is required to have the complete database at hand and initiate the huge process on analyzing the network from multiple aspects. The assembly of the database is the key for the analysis. This analysis is of course of huge help when a new version of the same database is created, with new data or information and then comparisons reveals the insights which can otherwise be never detectable. Having meticulously assembled the different databases from multiple sources, the work achieved in this thesis now allows to easily complying with the up to date information on the human proteins, emphasizing on the human platelets.

The complete human proteome with interactome information was compiled which revealed 54,218 simple interactions with over 4400 phosphorylation events between 10,915 human proteins. Additionally 135 dephosphorylation events were included into this interactome. The table 2 below depicts the numbers showing the increase of interactome information when compared with the previous version from us (*Dittrich, Birschmann et al. 2008*).

The comparison on the older to the latest version of PlateletWeb is important as it reveals the amount of data that was explored for this study along with the increased improvements in the platelet resources. The source of information was increased from a total of 6 proteomic studies to over 15 studies. Additionally, the data from the public databases like HPRD (*Keshava Prasad, Goel et al. 2009*), GPMDB and Uniprot, the transcriptomic database, SAGE (*Dittrich, Birschmann et al. 2006*) were taken and put together for the PlateletWeb. This would help to realize each of the platelet protein with its source as in which studies the protein is discovered as platelet, and in which databases this was as well defined. This complete analysis will help to classically consider the genuineness of the platelet protein.

The number of interactions of the platelet proteins was increased from 2839 to an impeccable number of 58758 making it more than 20 times bigger interactome network. This was achieved by considering all the well-known interactions in the literature and the databases. Similarly, the number of proteins itself was increased from 3093 to 19813, a factor of 6. This huge numbers indicate a much bigger network of proteins with much improved dataset when compared to the previous version.

This improved new version of PlateletWeb has a huge number phosphorylations and dephosphorylations, including the sites, the events and the kinase predictions in some cases. Additionally, each platelet protein was check to see if this is a phosphorylated kinase or protein alongside the complete protein information. This resource thus provides complete information on platelet proteins. The table 2 showcases the numbers that were increased form the older version to the present version of PlateletWeb database.

Table 2 - Human Interactome statistics:

Study/Database	***PlateletWeb 2008***	***PlateletWeb 2012***
Total interactions	2839	58758
Total proteins	3093	19813
Platelet proteins	3093	5025
Phosphorylation and Dephosphorylation events	-	6243

Study/Database	***PlateletWeb 2008***	***PlateletWeb 2012***
Total phosphorylation sites	-	73734
Total dephosphorylation events	-	171
Platelet phosphorylation events	-	28800
Platelet dephosphorylation events	-	99
Total phosphorylated proteins	-	10441
Total phosphorylated kinases	-	461
Platelet phosphorylated proteins	-	3532
Platelet phosphorylated kinases	-	216

The interactome was then used to check for its characteristics, and was also compared with the previous version of the platelet interactome network. This revealed an incredible increase in the platelet interactome (84.34%) which in turn enables to consider a high number of proteins for the statistical analysis.

The size of the largest connected component has changed from 1365 (70.65%) to 3060 (84.34%). This suggests the extremely good connectivity of the proteins with in the network. The network regulation can be investigated in detail because of this increase in the interaction and phosphorylation data.

The number of singletons is another important aspect, which describes the reliability of this network. The number of singletons remained almost constant. Considering the fact of the increased number of platelet proteins, these singletons signify the extremely good connectivity of the platelet network. The increase in connectivity also influences the average degree of nodes, which specifies the average number of interactions of each protein. This was improved from 2.951 to 6.771, a pretty higher connectivity of the platelet network. Also, the maximum path length has decreased from 15 to 10 indicating a tight moduled network.

Table 3 - Platelet Interactome Statistics

Network Parameter	***Interactome Dittrich2008***	***PlateletWeb2012***
Number of all nodes	1932	3628
Number of Interactions	2851	13652
Number of components	532	536
Size of largest component	1365	3060
Size of largest component (%)	70.65%	84.34%
Singletons	503	510
Singletons (%)	26.04%	14.06%
Average degree	2.951	6.771
Characteristic path length	4.972	3.917
Maximal path length	15	10

The proteins were then separated according to the residue and the numbers provide an overview of the distribution of Serine-and Threonine-phosphorylation sites in human platelets. Serine phosphorylations are predominant, followed by threonine and tyrosine phosphorylated residues. This tendency is found generally in all phosphorylation sites and specifically also in literature platelet sites. The experimentally-validated sites have a comparably low percentage of tyrosine phosphorylations (4.0%), which might be due to the lack of tyrosine-specific antibodies. According to previous studies, tyrosine phosphorylations are often found on less abundant proteins and their detection is complicated due to the lower stability the phosphorylated tyrosine (*Olsen, Blagoev et al. 2006*).

Table 4 - Phosphorylation sites

	S	***T***	***Y***
Literature all sites	45195	15249	13277
Literature platelet sites	16469	6180	6138
Experimental all / platelet sites	441	72	21

526 kinases were identified in the human proteome of which 229 were identified as platelet kinases. Additionally, the kinases substrate relations were calculated to determine how many platelet substrates have associated platelet and non-platelet kinases. When introducing a limitation of platelet kinases acting on platelet substrates, then the platelet kinases were reduced to 162 and the platelet substrates to 740. Association with a platelet kinase increases the chances of the phosphorylation to take place in platelets under normal biological conditions. Therefore, these phosphorylation events give strong indication for signaling regulation in the platelet when there is no experimental evidence available in literature for this cell type.

Table 5 – Kinase substrate associations; Kinase acting on substrates

	PlateletWeb 2012

	PlateletWeb 2012
Kinases on All substrates	329
Kinases on Platelet substrates	268
Platelet Kinases on All substrates	176
Platelet Kinases on Platelet substrates	162

Table 6 – Kinase substrate associations; Substrates targeted by kinases

	PlateletWeb 2012
Substrates targeted by Kinases	1801
Platelet Substrates targeted by Kinases	810
All Substrates targeted by Platelet Kinases	1614
Platelet Substrates targeted by Platelet Kinases	740

The Gene Ontology information is taken from the GO Consortium and all the human proteins were annotated in the *PlateletWeb*. The following table provides the statistics of Gene Ontology terms associated with the platelet proteins.

Table 7 –Gene Ontology information associated with Platelet proteins

	PlateletWeb 2012
Total GO Terms with all human proteins annotated in *PlateletWeb*	13736

	PlateletWeb 2012
Total GO Annotations with platelet proteins	10148
Total proteins associated with GO Term	17546
Platelet proteins associated with GO Term	4728

Along with the functionality of each platelet protein, relevant information was added about its drug associations. Drug modulation is an important aspect in cell biology because it allows controlled tuning of various cell processes and often represents a first step in development of new therapies. The drug information is taken form DrugBank, and then mapped with the platelet proteins in the PlateletWeb resource, which revealed a high number of platelet drug targets. Almost half of the known human drugs (2706 out of 4311 drugs) are targeting platelet proteins. The high amount of platelet drug targets indicates that drug discovery is focused on many proteins in the platelet cell and underlines the importance of platelets in clinical research.

Table 8 – Drug Targets

	PlateletWeb 2012
Total Drug Targets	2106
Platelet Drug Targets	950
Total Drugs	4311
Drugs acting on platelet proteins	2706

3.1.3 Knowledgebase and the flow of information

The platelet data is accumulated from various sources using the techniques of data mining and manual curation. In order to investigate the biological insights of this data, a resource was created which helped to understand platelet signaling pathways. This resource is further extended to create a website (knowledgebase) for platelets as a platelet systems biological platform which was termed *PlateletWeb*. This knowledgebase combines and visualizes the most recent proteomic studies investigating platelet proteins and provides basis to a more comprehensive analysis of platelet-specific functions. Based on data from our resource gave a complete overview of platelet signaling, interactions, phosphorylation events and identified enrichment of platelet-specific pathways and functions. The kinome and phosphoproteome were analyzed, followed by the introduction of drug options and network modulation. Furthermore, the physical information of the protein along with the Gene Ontology and KEGG pathway information was also visualized and made available in the knowledgebase. The platform allows a comprehensive and detailed analysis of the platelet not only on a single protein level but also on the scale of network regulation and functional association of signaling components. Alongside, the access of the information contained in the resource is made easy by creating multiple search functions which would help to extract the information easily and quickly from the knowledgebase.

Assembling the database itself is one main aspect; however it is not finished without creating a proper graphical user interface to retrieve the right amount of information. The accessing of the information would require sophisticated website development to give the ease of use. This would make the database a user friendly resource which can be easily accessible from anywhere around the world. A first hand website is created for the previous version (*Dittrich, Birschmann et al. 2008*), which was extended to add the new information with a high significance on the platelet proteins.

A detailed tutorial on the usage of the platform is presented in the supplemental Appendix 1 (*PlateletWeb* user guide).

The following flowchart depicts the way the key aspects of PlateletWeb knowledgebase and the way it helps to access the right data. The knowledgebase design is biased towards the end user’s perspective. For instance, it is possible to get the complete details of the protein starting from the interactions, its phosphorylations, its site specific information, the physical properties etc., by just entering the name of the protein. Additionally, the similar names, which would give another chance to recheck the protein is also considered. Some proteins have alternative names and this presentation of similar proteins would help in order to get the information for the right protein. The complete protein description page, containing multiple sections, each of which, concentrating on a particular information of the protein can also be retrieved. This is a simple search which provides the ease of use and gets the information about the protein instantly.

It is also possible to concentrate on a group of proteins with similar functionality or to extract a group of proteins which are, for instance, phosphorylating on the tyrosine residue. There is multiple simplified advanced search criteria were defined with in the knowledgebase which would give access to retrieve the set of proteins with similar classification. These proteins can be either searched from the point of view of the drug targets or the containing similar gene ontology term or at the lowest level having similar physical properties. A protein or group of proteins with similar attributes is retrieved in such cases and this can then be also visualized. For example, in the advanced search it is possible to extract platelet proteins with multiple similar features: having the same function (e.g. hemostasis) and protein domain (e.g. transmembrane domain TM) and phosphorylated on a tyrosine residue. This particular search yields three proteins with the predefined criteria: GP1BA, ITGB3 and PECAM-1. All three proteins are well-known platelet signaling modulators, playing a role in platelet activation using SH2 domain binding.

Additionally, the information obtained at any instance, either at the level of similar proteins or the phosphorylations of a particular protein or even the description of the protein, they all can be printed. Apart from this, the download options enable to download specially the information about the sequences and the subnetworks that are visualized. The flow chart gives the pictorial representation of the complete work flow of the website.

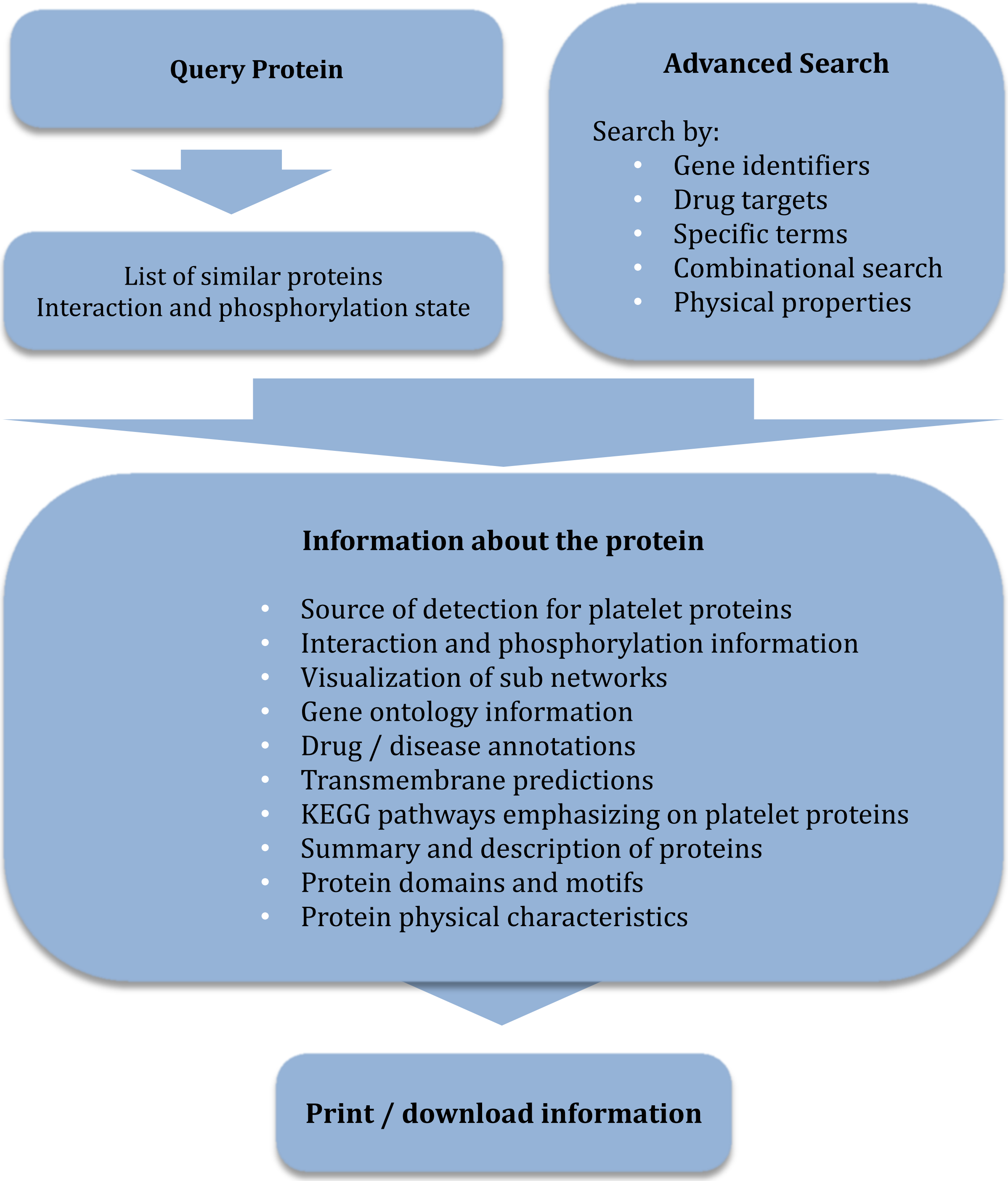

Query Protein
List of similar proteins
Interaction and phosphorylation state
Advanced Search
Search by:
Gene identifiers
Drug targets
Specific terms
Combinational search
Physical properties
Information about the protein
Source of detection for platelet proteins
Interaction and phosphorylation information
Visualization of sub networks
Gene ontology information
Drug / disease annotations
Transmembrane predictions
KEGG pathways emphasizing on platelet proteins
Summary and description of proteins
Protein domains and motifs
Protein physical characteristics
Print / download information

Figure 2 – Flowchart of the PlateletWeb workbench.

The query can be provided using simple search or advanced search. This provides the complete information about the protein(s) and it is also possible to download or print the result.

The database is very useful to investigated interactions in the platelet. By choosing the protein name (e.g. VASP) an overview of the protein description, physical and biological properties and phosphorylations is revealed. In contrast, to find the interaction partners such as Abi1, a network view of the neighboring proteins can be obtained. This view is extended by introducing a new option to extract a set of proteins with their network context, thus allowing a detailed investigation of phosphorylations and interactions of specific protein datasets.

3.1.4 Subnetwork extraction

In order to gain the perspective of the proteins in the network format, the visualization plays a crucial role. This would give the required perspective and orientation on the proteins that are of interest. In order to visualize the proteins in the network format, an additional feature was added to the PlateletWeb resource. Subnetworks can be extracted from the knowledgebase for proteins identified in proteomic analysis or in a given sample of proteins. This can help to visualize and understand the proteins in their network context and further emphasis can be given onto the signaling modules. The combination of proteomic and transcriptomic information of the platelet proteins, along with the phosphorylation, dephosphorylation and interaction data together with the drug information and the site- specific information would give a maximum insight of the signaling process within the module. Additionally, the differentiation between kinases, substrates, platelet proteins, non-platelet proteins and drugs would help to analyze the subnetwork module in a much more systematic way. The resulting network can be easily saved onto the computer and visualized in vector graphics or can be further imported and investigated using the network analysis software, Cytoscape.

Case study: LXR and PECAM1 interactions

The PlateletWeb resource can be used to investigate the pathways and all the interactions of any platelet proteins of interest. Newly identified antithrombotic targets such as LXR (*Spyridon, Moraes et al. 2011*) or the signaling modulator PECAM1 (*Moraes, Barrett et al. 2010*) can be easily investigated in the context of interacting partners and phosphorylation events.

LXR interactions:

The central network around the two isoforms of the liver X receptor (LXRα and LXRβ), presents all interactions and known phosphorylation events obtained using the subnetwork extraction feature of *PlateletWeb*. The LXRβ isoform has been detected in platelets and it is therefore visualized in yellow. Phosphorylated platelet proteins are presented in red. The interaction network reveals a kinase-substrate relationship between the LXRα and the casein kinase CSNK2A1. Furthermore, the PlateletWeb knowledge base presents options to explore the antithrombotic effects around the LXRβ. By first extracting the neighboring interactions and then focusing on the GP6 receptor downstream signaling implied to be associated with LXR binding effects observed in platelets (*Spyridon, Moraes et al. 2011*), the connection between the two can be thoroughly examined. For instance, LXRβ is directly interacting with the ryanodine receptor RXRA in platelets, which is associated with the Src kinase. This kinase along with the kinases SYK, LYN and FYN is one of the major activatory kinases in human platelet signaling (*Varga-Szabo, Pleines et al. 2008*). It is also very closely associated with the GP6 receptor and the LAT protein. Thus, by using interaction and phosphorylation information available in PlateletWeb complex network associations can be investigated and visualized.

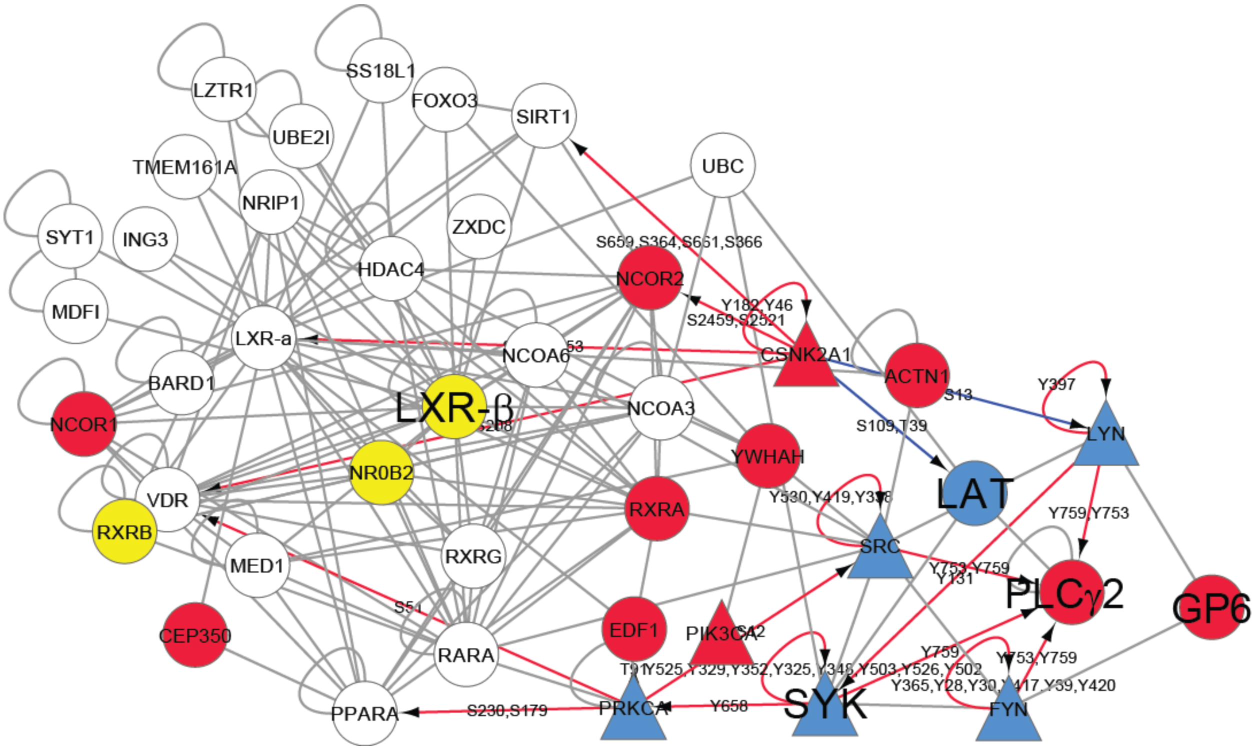


Figure 3 - Interaction and phosphorylation network of LXR

Visualization of this LXR interaction network is created by integrating information from multiple sources. The circles indicate the proteins, while the triangles indicate the kinases. The interactions in the network are depicted using the gray lines and the phosphorylations and dephosphorylations using the green lines. Phosphorylations are in turn specified according to their source of detection: red arrows indicate phosphorylations reported from human cells (HPRD), blue arrows are used in the cases where a kinase prediction is assigned to an experimentally-validated phosphorylation site. The protein nodes are colored according to the source of phosphorylation (red indicating that the protein is phosphorylated in human cells and blue indicating that the protein is phosphorylated in platelets. The yellow color specifies a non-phosphorylated platelet protein. The phosphorylation site is presented on each directed edge, providing further information of the kinase phosphorylating on the site.

PECAM1 interactions

One of the essential mechanisms of integrin outside-in signaling is integrin clustering, which then leads to dephosphorylation of Src at Y530 and subsequent autophosphorylation at Y419. Pecam-1 acts as a positive regulator in integrin a2bb3 engagement, clustering and SHP-2 recruitment and activation (*Jackson, Kupcho et al. 1997; Jackson, Ward et al. 1997*). If SHP2-Pecam1 localization is absent, dephosphorylation of Src doesn't take place, which leads to defective downstream signaling and delay in cell spreading, clot retraction and focal adhesions. Phosphorylation of both ITIM-motifs of PECAM1 (Y663, Y686 = Y690, Y 713) is needed for recruitment and activation of SHP2 (= PTPN11) (*Newman and Newman 2003*). Example of the regulation of collagen stimulated platelets through PECAM-1. The modification of the interaction of PI3KR1 to LAT and GRB2-associated binding protein 1 (GAB1) (*Moraes, Barrett et al. 2010*) can be visualized in a network context additionally including SRC to examine phosphorylation events and identify PTPN11 (also known as SHP2) as key counter-player for dephosphorylation events.

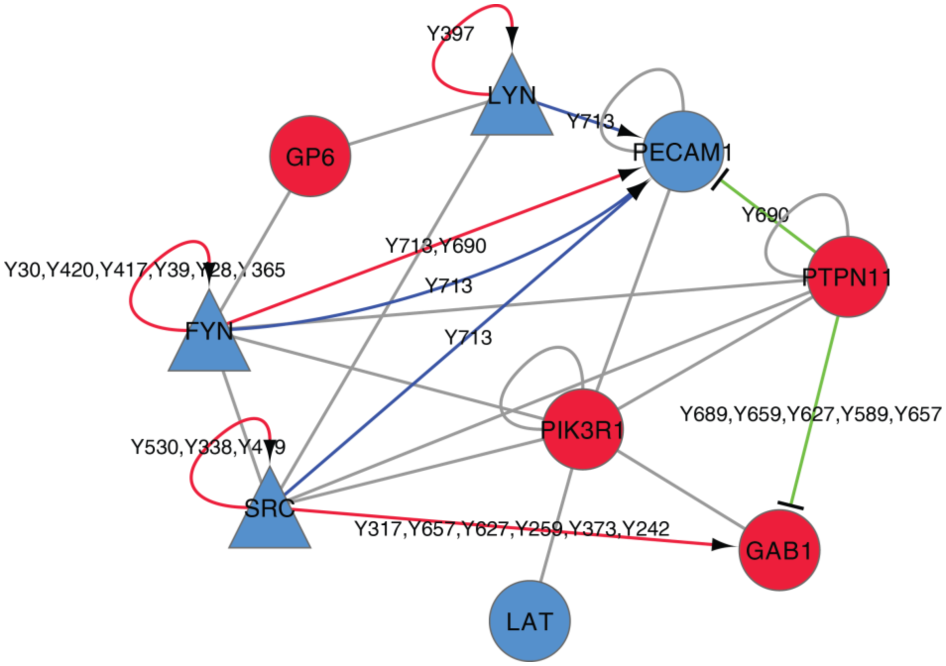


Figure 4 - Interaction and phosphorylation network of PECAM1

The interaction network of PECAM1 showing its interactions, phosphorylations along with the kinases and the proteins associated with it.

However, the subnetwork extraction is not restricted to a single protein. The following figure represents the visualization of the vWF signaling subnetwork extracted from the *PlateletWeb* knowledgebase which contains the list of platelet proteins, non-platelet proteins, kinases, substrates, interactions, phosphorylations, dephosphorylations, drugs, and drug targets. However, only a few drugs were shown in the picture for improved visualization and simplicity.

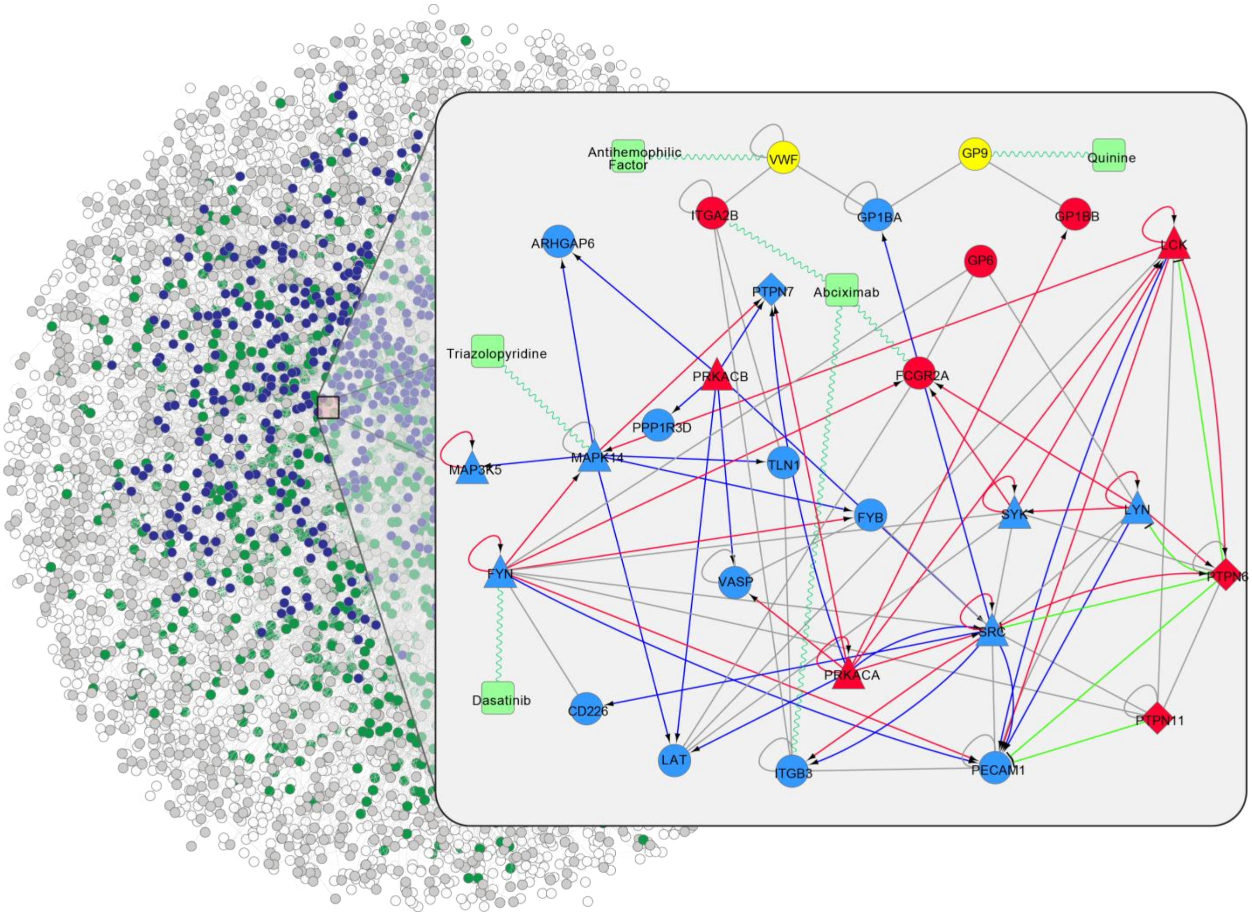


Figure 5 – Extraction of subnetwork from the complete human interactome

A subnetwork can be extracted from the human interactome including the protein information, the phosphorylations, the interactions and the drug targets.

3.1.5 SAGE/Proteome Distribution

Considering all the SAGE (*Dittrich, Birschmann et al. 2006*) and proteomic data for the platelet proteins, an enrichment analysis was performed. The main challenge is to find and integrate the vast amount of reliable data which is available from multiple resources. The figure below shows the proteins detected on the SAGE and proteome level in biological processes of the platelet proteins. The analysis reveals that the term "translation" contains the highest percentage of SAGE proteins. This is partly carry-over effect from Megakaryocytes, however, there is also translation in platelets, in particular upon activation (as shown in the example by the Weyrich group (*Weyrich, Schwertz et al. 2009*)). An extremely pure preparation of platelets (*Dittrich, Birschmann et al. 2006; Dittrich, Strassberger et al. 2010*) was used for library preparation.

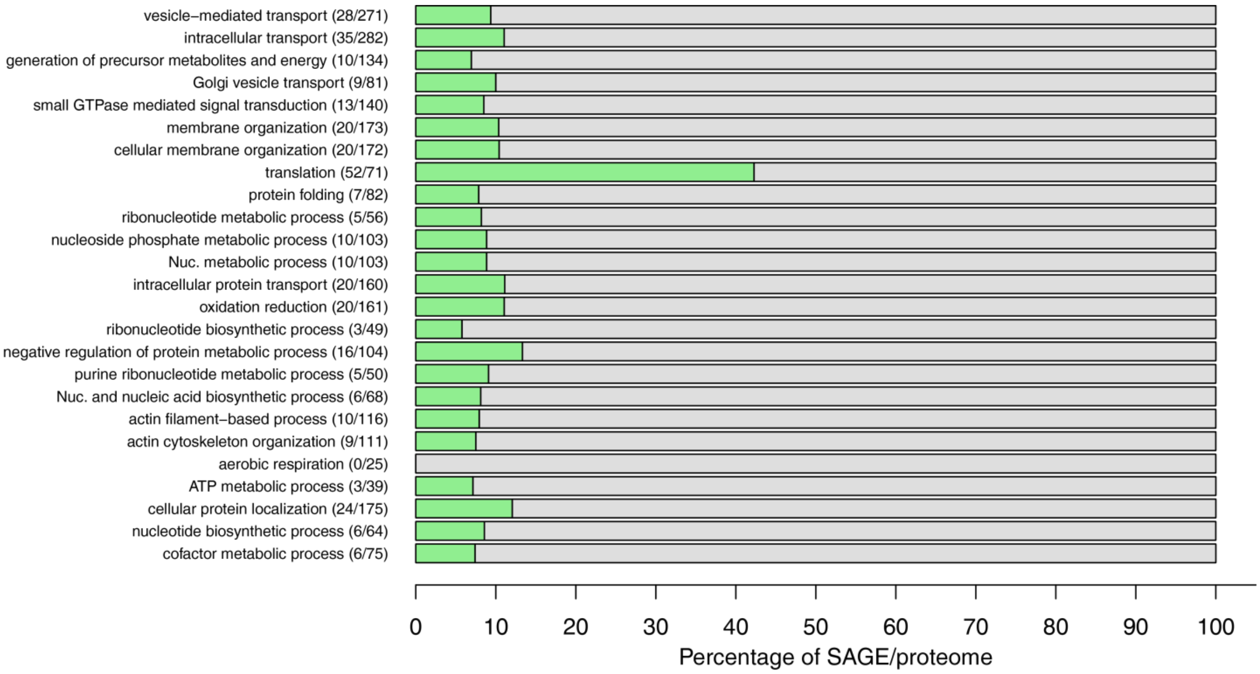


Figure 6 - Distribution of platelet proteins according to detection level

3.1.6 Platelet Enrichment

Biological processes were tested for enrichment in platelets. The top most significantly enriched specific terms were plotted against the negative log10 of the P value and colored according to their general functional category. Number of platelet proteins and total proteins for each term are given in parentheses. There is enrichment of terms related to transport, membrane organization, actin-filament based processes, and GTPase-mediated signal transduction.

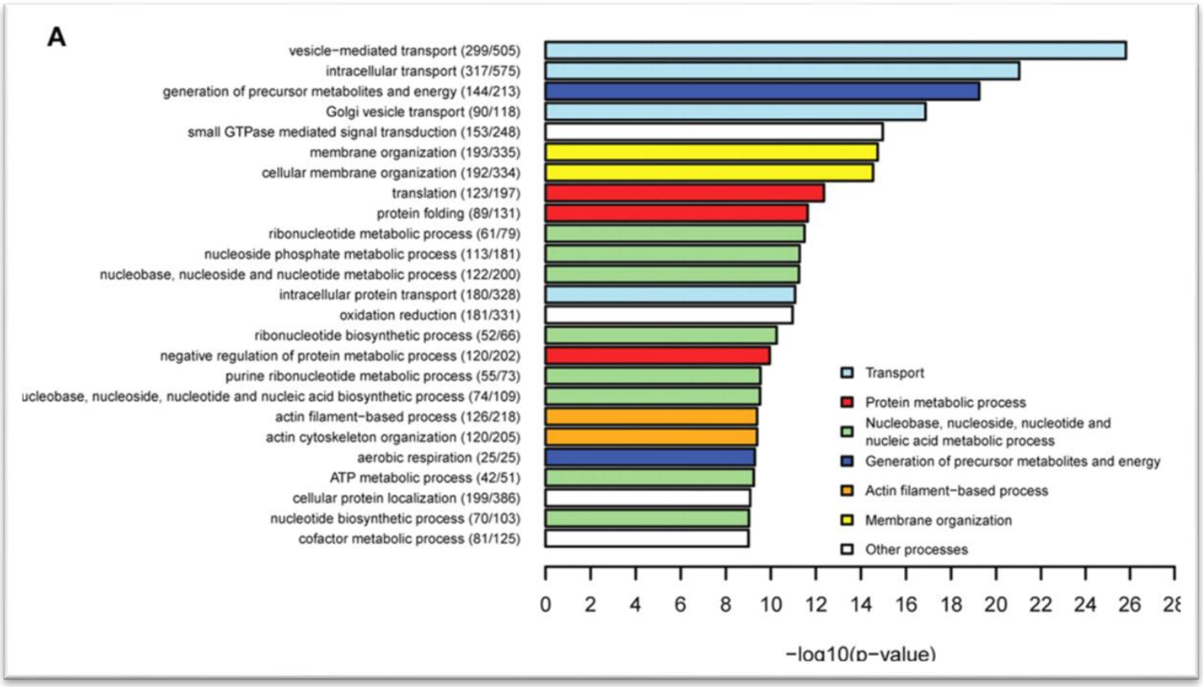


Figure 7 - Platelet enrichment: Functional analysis of Biological Process

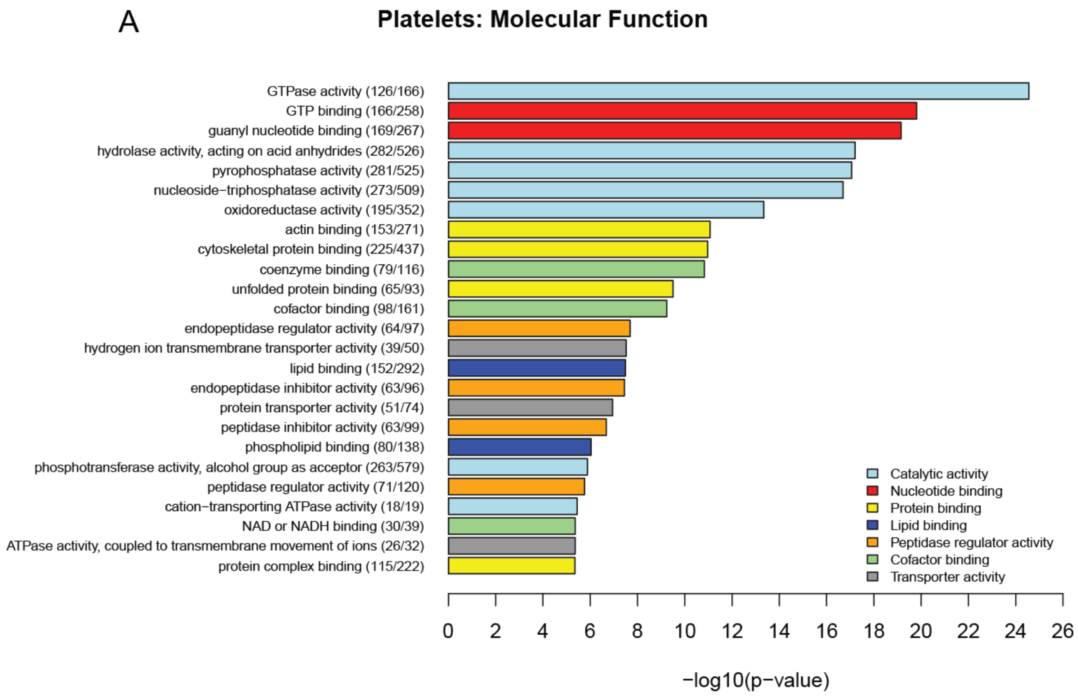


Figure 8 - Platelet enrichment: Functional analysis of Molecular Function

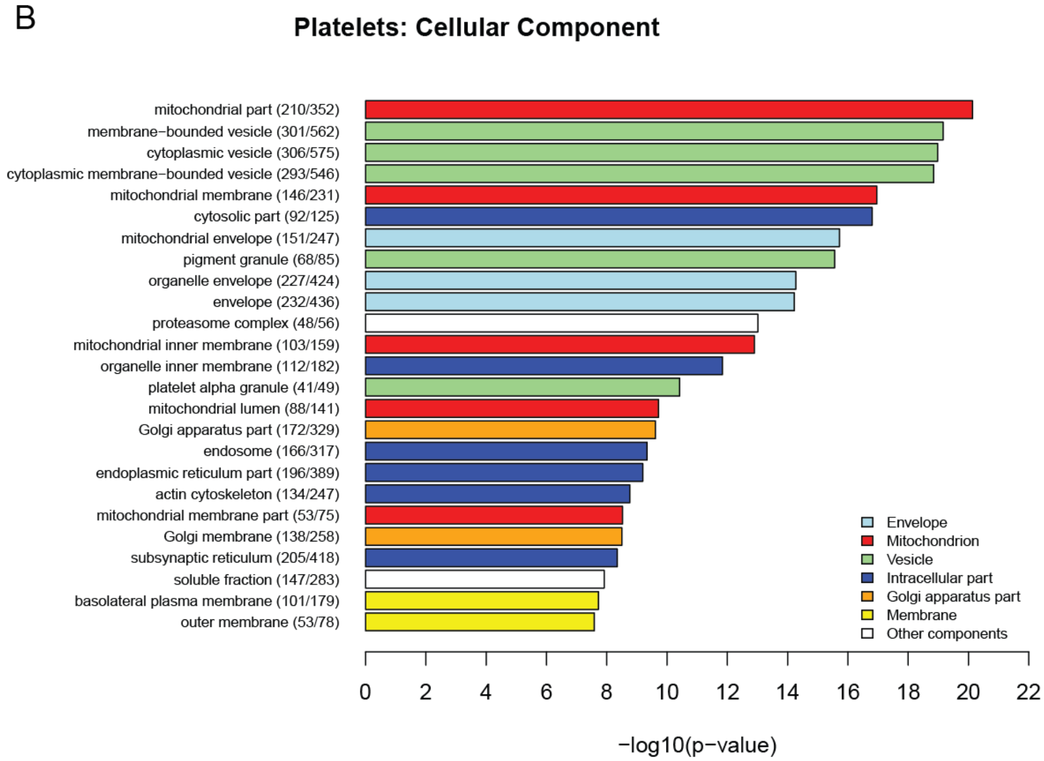


Figure 9 - Platelet enrichment: Functional analysis of Cellular Component

Functional characteristics of platelet genes and gene products are stored in the knowledge base, including the GO categories. This allows a functional characterization of the platelet network. Comparison of the GO functional categories of the platelet interactome with those of the entire human interactome network indicated a significant enrichment of platelet relevant biologic processes such as "vesicle-mediated transport" (P value = 1.54×10^{-26}), "small GTPase mediated signal transduction" (P value = 1.08×10^{-15}), and "actin cytoskeleton organization" (P value = 4.09×10^{-10}) (Figure 7). These processes are well established in the platelet as they are needed for platelet exocytosis, cytoskeleton organization and shape change. Similarly, enriched molecular functions mediate key platelet processes such as "actin binding" (P value = 8.51×10^{-12}) and "cytoskeletal protein binding" (P value = 1.09×10^{-11}) (supplemental Figure 8A), while cellular components were enriched for mitochondrial compartments and membrane components (supplemental Figure 8B).

3.1.7 Platelet-specific kinase tree reveals distribution and enrichment of tyrosine kinase substrates in human platelets

Site-specific phosphorylations from public databases were extracted to incorporate the central phosphorylation signaling network into the *PlateletWeb* resource. This was mainly based on experimental data of human cells and tissue expressions published in literature (*Hornbeck, Chabra et al. 2004; Keshava Prasad, Goel et al. 2009*). The extracted human phosphorylations consist of 10,441 human proteins with 73,734 phosphorylation sites. From them, 28,800 of the phosphorylations account for platelet phosphorylation sites, which represent around 39% of the total phosphorylation sites. These literature-derived data were complemented by a set of 533 phosphosites experimentally measured in human platelets (*Zahedi, Lewandrowski et al. 2008*), of which 16 sites have not yet been described in the human proteome.

Data on both human and platelet phosphorylation sites were further distributed according to the residue and this showed a majority of sites (over 50% in both cases) are on a serine residue and an almost equal amounts of threonine and tyrosine phosphorylations (around 22% each, in both cases). However, in contrast, the experimentally validated platelet phosphosites contain a substantially lower share of around 4% tyrosine phosphorylations combined with higher percentage (around 80%) serine phosphorylations.

By supplying information about the responsible kinase, phosphorylation sites can be analyzed in the network context. The main theme for this would be the kinase-substrate relationship. In this regard, the focus is mainly on the phosphorylated platelet proteins associated with a kinase. These represented only 23% (814) of all phosphorylated platelet proteins (3,532). On the other hand, 3,080 distinct phosphorylation sites have the associated kinase. However, kinase data for experimentally validated phosphorylations was available for only 69 phosphosites. A novel network-based algorithm was used to predict potential kinases for these sites (*Linding, Jensen et al. 2007; Miller, Jensen et al. 2008*), which resulted in kinase predictions for a further 436 sites yielding a total kinase annotation for 505 (94.5%) phosphosites. When introduced into the entire platelet phosphoproteome, these predictions contribute to 16% of all modification events with available kinase or phosphatase information

The entire platelet proteome dataset contains 229 kinases (43.5% of 526 total human kinases), 162 (70.7%) of which have well described substrates in the platelet proteome. Nearly all (216, 94%) have documented phosphorylation sites. The kinase families TK, CMGC and AGC contain the highest percentage of platelet kinases. The number of kinases belonging to each kinase group is separated into numbers for platelet kinases and non-platelet kinases. Then, the percentage of platelet and non-platelet kinases in the respective groups is calculated. A total of 218 platelet kinases and 295 non-platelet human kinases are considered.

Dephosphorylation is achieved by 73 platelet phosphatases (38.2 % of 191 total human phosphatases) of which 24 have characterized substrates.

Table 9 - Platelet kinome in relation to all human kinases

Kinase family	*Platelet Kinase*	*Non-platelet kinases*	*Platelet kinases (in %)*	*Non-platelet kinases (in %)*
AGC	34	29	15.60	9.83
Atypical	12	26	5.50	8.81
CAMK	22	52	10.09	17.63
CK1	5	7	2.29	2.37
CMGC	35	28	16.06	9.49
Other	21	57	9.63	19.32
RGC	0	5	0.00	1.69
STE	27	20	12.39	6.78

Kinase family	*Platelet Kinase*	*Non-platelet kinases*	*Platelet kinases (in %)*	*Non-platelet kinases (in %)*
TK	51	39	23.39	13.22
TKL	11	32	5.05	10.85

All platelet kinases were mapped onto the human kinome tree (*Miller, Jensen et al. 2008*) in order to analyze them and their substrates in a phylogenetic context.

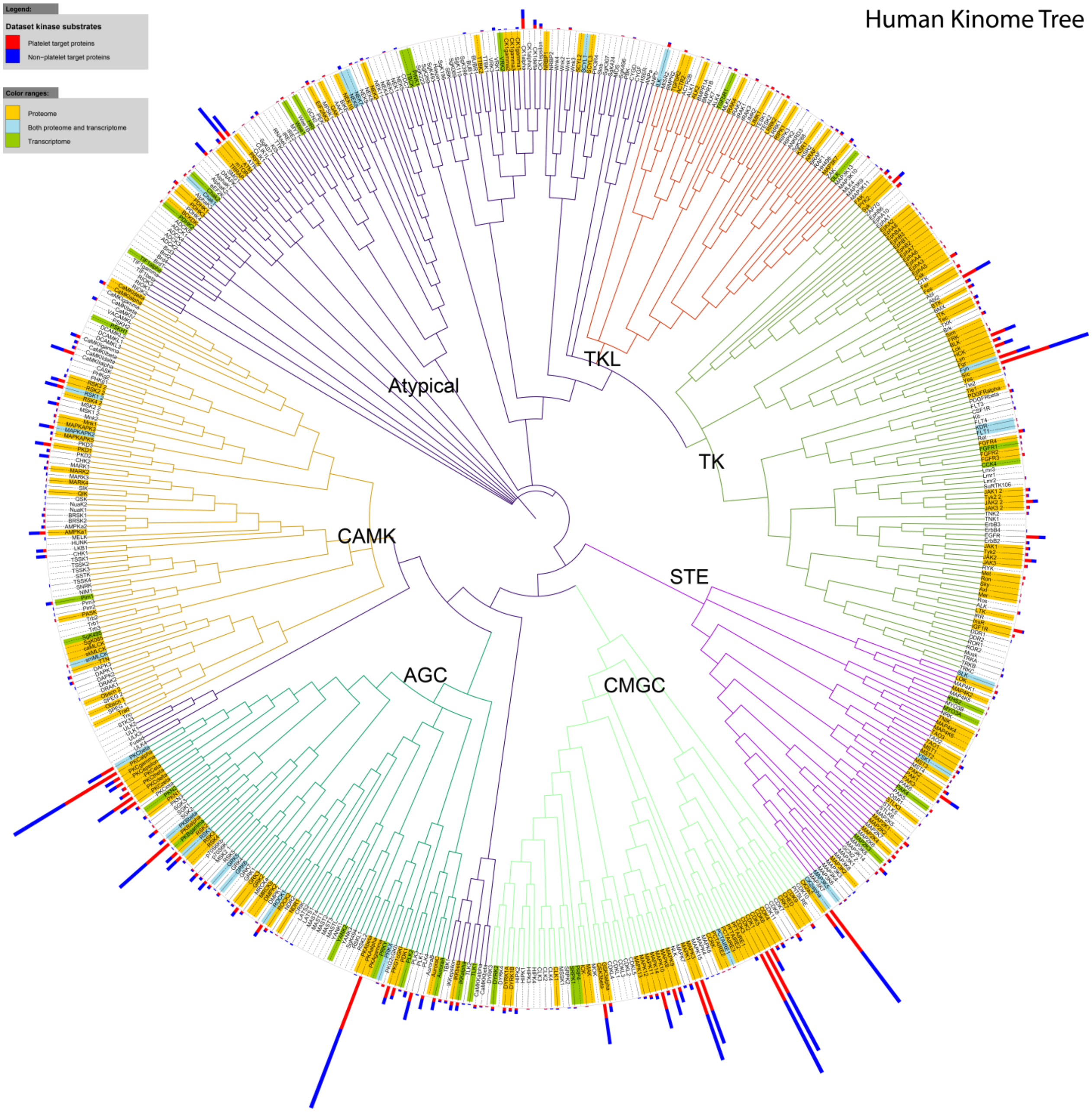

Human Kinome Tree
Legend:
Dataset kinase substrates
🟥 Platelet target proteins
🟦 Non-platelet target proteins
Color ranges:
🟨 Proteome
Both proteome and transcriptome
🟩 Transcriptome
Atypical
TKL
TK
CAMK
STE
AGC
CMGC

Figure 10 – Human Kinome Tree

Most of the platelet signaling pathways relies on kinases, which allow phosphorylation of many proteins in the platelet after stimulation. The assembled platelet interactome consists of 229 platelet kinases from overall 513 human kinases. Mapping the platelet kinase 1and substrate data onto the phylogenetic human kinome tree we marked platelet kinases according to their level of platelet expression (proteome, transcriptome or both) and show the number of platelet and non-platelet substrates for each kinase as two-coloured bars next to its *name. Substrates information is presented as integrated into PlateletWeb from human phosphorylation databases. The overall distribution evidences abundance of PKC kinases and Src-family which are found in most of the platelet activation signaling pathways.*

Mapping the platelet kinase and substrate data onto the phylogenetic human kinome tree the platelet kinases according to their level of platelet expression (proteome, transcriptome or both) were marked and showed the number of platelet and non-platelet substrates for each kinase (data in the supplement). The majority of platelet tyrosine kinases were found on the proteome level. In order to get the detailed analysis, the tyrosine kinase subtree was extracted for detailed analysis from the main phylogenetic tree.

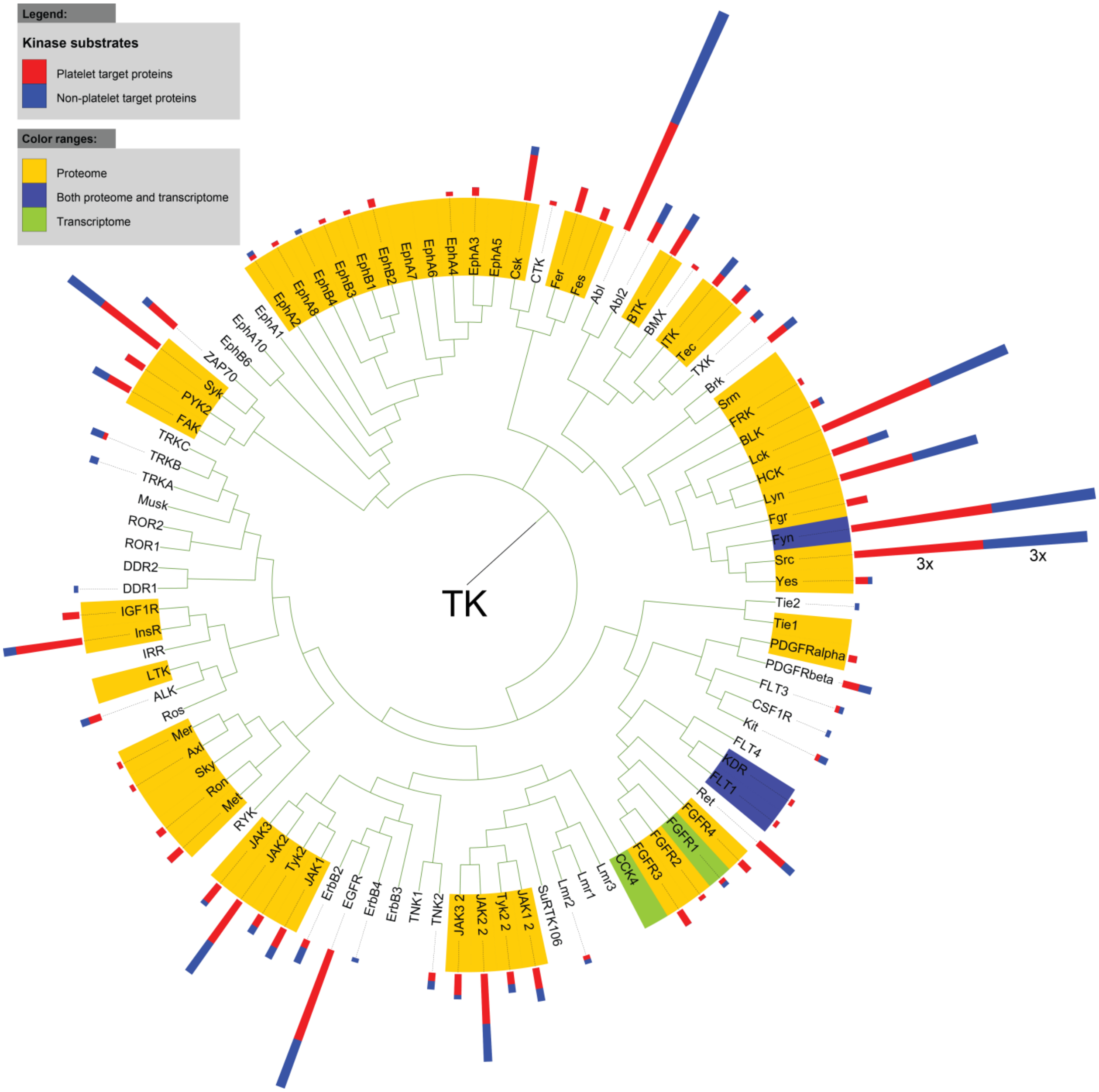


Figure 11 - Tyrosine Kinase tree

The Tyrosine kinase subtree reveals a strong representation of Src-family kinases, which have many characterized substrates in platelets and mediate platelet activation through numerous phosphorylation events.

In order to get the detailed overview of the kinase families (*Manning, Whyte et al. 2002*), an enrichment analysis was performed. The tables below show the results of the enrichment analysis.

Table 10 - Statistical analysis of kinase and kinase substrates' enrichment

Kinase family	*Platelet kinases*	*Non-platelet kinases*	*Total kinases*	*p-value*	*Corrected p-value*
TK	51	39	90	0.01	0.08
Other	21	57	78	0.01	0.08
CAMK	22	52	74	0.04	0.11
AGC	34	29	63	0.11	0.13
CMGC	35	28	63	0.06	0.11
STE	27	20	47	0.06	0.11
TKL	11	32	43	0.04	0.11
Atypical	12	26	38	0.23	0.26
CK1	5	7	12	1	1
RGC	0	5	5	0.08	0.11
All kinases	218	295	513		

Enrichment of kinases in platelets according to the phosphorylated residue (** Dual-specificity kinases are not represented, therefore the sum of the column doesn't equal the number shown)

Table 11 - Enrichment of kinases in platelets according to the phosphorylated residue

ST and Y Kinases	*Platelet kinases*	*Non-platelet kinases*	*Total kinases*	*p-value*	*Corrected p-value*
ST Kinases	136	215	351	0.29	0.29
Y Kinases	51	45	96	0.06	0.12
All ST and Y Kinases**	187	260	447		

Enrichment of kinase substrates in platelets according to the phosphorylating kinase family (* The substrates can be targeted by more than one kinase family, therefore the sum of the column doesn't equal the number of total substrates)

Table 12 - Enrichment of kinase substrates in platelets according to the phosphorylating kinase family

Kinase substrates	*Platelet substrates*	*Non-platelet substrates*	*Total substrates*	*p-value*	*Corrected p-value*
CMGC	363	463	826	0.67252	0.76
AGC	308	360	668	0.584902	0.75
TK	222	167	389	1.52E-05	**1.37E-04**
CAMK	89	138	227	0.118601	0.18
STE	74	52	126	0.002967	**0.01**
Other	60	107	167	0.027813	0.05
Atypical	47	90	137	0.02	0.045

Kinase substrates	*Platelet substrates*	*Non-platelet substrates*	*Total substrates*	*p-value*	*Corrected p-value*
TKL	35	21	56	0.009605	**0.03**
CK1	24	28	52	0.888071	0.89
All Substrates*	804	989	1793		
CMGC	363	463	826	0.67252	0.76

Enrichment of kinase substrates in platelets according to the type of phosphorylating kinase - (* the substrates can be targeted by more than one kinase family, therefore the sum of the column doesn't equal the number of total substrates)

Table 13 - Enrichment of kinase substrates in platelets according to the type of phosphorylating kinase

ST and Y substrates	*Platelet substrates*	*Non-platelet substrates*	*Total substrates*	*p-value*	*Corrected p-value*
All distinct substrates of ST Kinases	580	741	1321	0.609655	0.61
All distinct substrates of Y Kinases	222	167	389	1.52E-05	**3.04E-05**
All ST and Y substrates*	804	989	1793		

Although kinases of the tyrosine and AGC family are relatively abundant in human platelets, no significant overrepresentation of a particular kinase family could be detected after multiple testing correction. Analogously, the enrichment of platelet kinase substrates according to kinase family and residue was investigated. The analysis yielded a significant enrichment of platelet substrates of STE (*P* = 0.013), TKL (*P* = 0.03) and predominantly substrates of the TK kinase family (*P* value = 1.4×10^{-4}), also reflected on the residue level by a significant enrichment of tyrosine-specific kinase substrates. The STE family contains a broad repertoire of kinases important for the upstream signaling in the MAPK/ERK pathway (MAPKKK kinases). The TK group contains the main activatory kinases of the Src-family for which a large number of platelet substrates have been and they participate in processes such as collagen and vWF-induced platelet activation (*Gardiner, Arthur et al. 2010*).

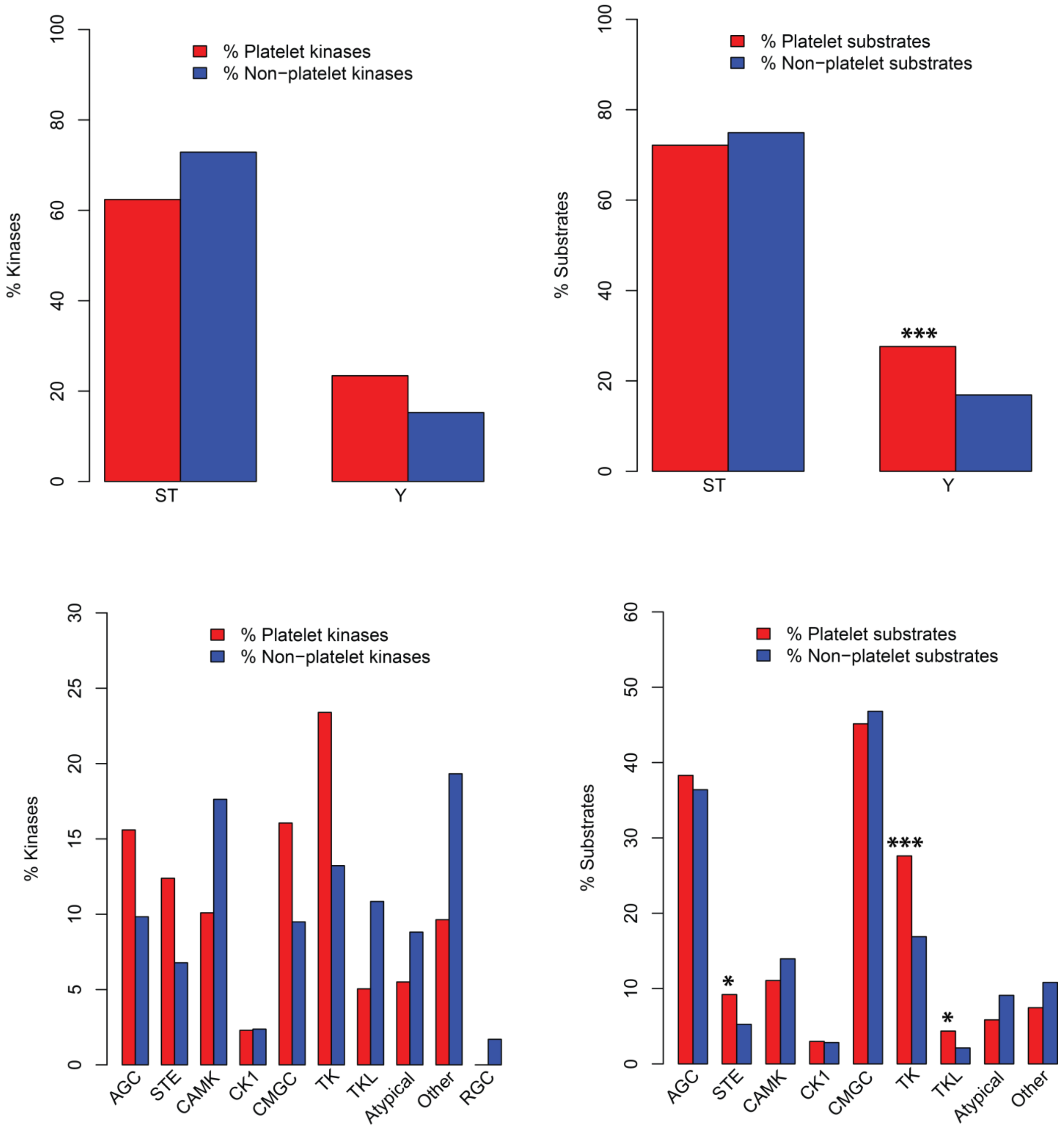


Figure 12 – Enrichment of kinases and substrates with respect to the residues and kinase families

3.1.8 Protein domains

Further analysis indicated that gated interaction mediating domains SH2 (supplemental Figure 5) are overrepresented in platelets, consistent with their role as counterparts of tyrosine kinases through recognition of tyrosine phosphorylated residues (*Liu, Jablonowski et al. 2006*). The most often occurring protein domains in platelets are listed. The y-axis depicts the –log10 of the *P* value (Fisher test). The top 10 significantly enriched domains include RAB domain, Serine threonine kinase and Tyrosine kinase domain. Furthermore, the gated interaction mediating domains SH2 and SH3 are overrepresented due to their crucial role in platelet activation.

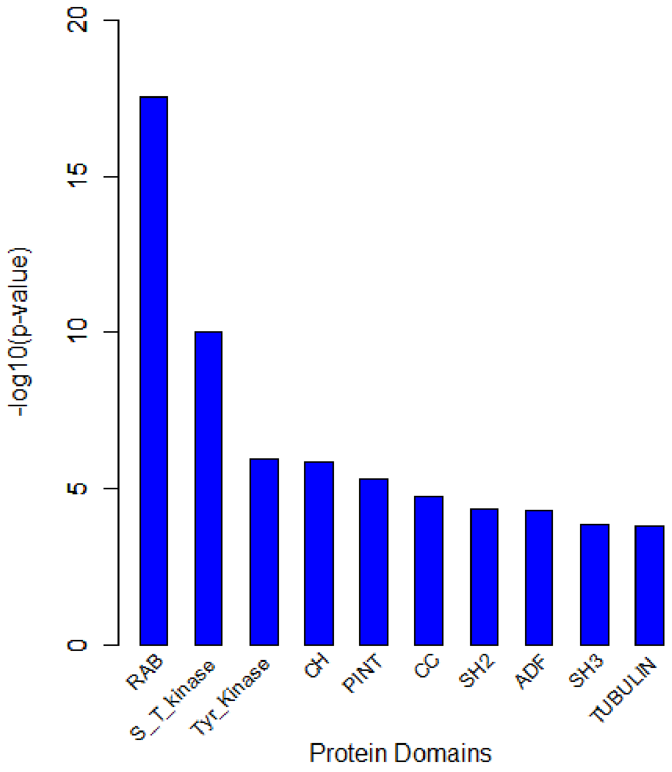


Figure 13 – Protein domains

Enrichment of protein domains and their distribution

Analyzing the building blocks of the platelet signaling network in terms of domain architecture, many of the most frequent domains among hub proteins (platelet proteins with more than 50 platelet interactions) are mediators for protein-protein interactions including adaptor domains. These include the gated interaction mediating domains SH2, SH3 as well as kinase domains which are crucial for signaling and dynamic switches in complex formation of the highly connected central phosphorylation network. Domains found in platelet proteins with the highest number of interactions (a total of 27 platelet proteins with over 50 platelet interactions, some proteins containing more than 1 listed domains). The number of proteins containing the respective domain is shown.

Figure 14 – The number of proteins with over 50 interactions contained in the domain

Number of proteins with over 50 interactions	***Domain Name***	***Domain Description***
7	SH2	Src homology 2 domains
7	S_T_kinase	Serine_Threonine protein kinases, catalytic domain
6	SH3	Src homology 3 domains
5	CC	Coiled Coil
5	Tyr_Kinase	Tyrosine kinase domain
2	ACTIN	Actin
2	C1	Protein kinase C conserved region 1 (C1) domains (Cysteine-rich domains)

Number of proteins with over 50 interactions	***Domain Name***	***Domain Description***
2	C2	Protein kinase C conserved region 2 (CalB)
2	TM	Transmembrane Domain
1	ARM	Armadillo/beta-catenin-like repeats
1	B41	Band 41 homologues
1	CASc	Caspase, interleukin-1 beta converting enzyme (ICE) homologues
1	GS	GS motif
1	LIM	Zinc-binding domain present in Lin-11, Isl-1, Mec-3
1	MH1	MAD homology 1
1	MH2	MAD homology 2
1	PH	Pleckstrin homology domain
1	RHO	Rho (Ras homology) subfamily of Ras-like small GTPases
1	RHOGAP	GTPase-activator protein for Rho-like GTPases
1	SP	Signal Peptide
1	Tyr_Phos	Protein tyrosine phosphatase, catalytic domain

Number of proteins with over 50 interactions	***Domain Name***	***Domain Description***
1	UBQ	Ubiquitin homologues

The results conclude multiple facts considering PlateletWeb as a base. The comprehensive dataset which was assembled contained various sources with high manual curation information. Most of the proteins detected exclusively on the proteome level make the dataset more reliable. The proteins identified on the transcriptome level, also contributed equally to the *PlateletWeb* repository. The fact that platelets don't have nucleus means that they can't perform the transcription, however translation is definitely possible if mRNA is left from the megakaryocytes from which the platelets are derived. Platelets contain a pool of mRNA which can be spliced and translated in a signal-dependent manner (*Denis, Tolley et al. 2005; Dittrich, Birschmann et al. 2006; Schwertz, Tolley et al. 2006; Rowley, Oler et al. 2011*). Large-scale proteomic data is crucial for understanding the systems biological background of cellular processes as well as for identifying potential new biomarkers in human diseases as recently shown for atherothrombosis (*Tunon, Martin-Ventura et al. 2010; Dittrich, Birschmann et al. 2011*) .

Interestingly, there are a lot more phosphorylation events than dephosphorylation events available from HPRD. This can be explained by the fact that phosphorylation triggers most of the activatory and inhibitory pathways in platelets and a tight regulation is needed to ensure signaling events take place in an organized and strictly defined manner. Therefore, a high number of kinases are needed to fine-tune platelet's responses.

The PlateletWeb database is flexible to include and extend any newly available data or change in the present information. The extensive usage of the Perl scripts and the MySQL database help the easy transformation and inclusion of any additional information into the resource. Additionally, the website design is also easily extensible and scalable to cope up with huge amount of information. Both the database and the resource allow the additional extensions which might be required when completely new functionality has to be added. Apart from the additional features, there is a huge possibility to bridge multiple species together. For instance, a protein is found in the human proteome is identified as a platelet protein. It is also possible to scale this information on multiple species for example on a mouse and see the further information of the same protein in the mouse proteome and get the detailed analysis of it. This example shows the capability of this resource when an exponential growth of the database with additional information at sight. .

The platelet interactome statistics revealed few key notes when compared with the previous version. The increase in the largest component, (from 70% to 84%) showcases the fact that the proteins are connected between themselves leaving very few singletons behind (decreased from 26.04% to 14.06%). All the enrichment analysis performed was based on the connecting graphs, and this implicates that a reasonable amount of proteins were considered for the enrichment analysis that was performed. This higher number of proteins participating in the enrichment analysis provides with more concrete results. Furthermore, the average degree of the network has a whooping increase from 2.951 to 6.771 illustrates the higher connectivity possessing in the network. This was possible because of the increased number of interactions included with in the network.

The identification and assembly of the kinases into the PlateletWeb provided a strong base for understanding the platelets in more detail. A phosphorylation network consisting of all the kinases and networks was also been assembled for further analysis. A comprehensive list of human kinases extracted from Manning et al (*Manning, Whyte et al. 2002*) and later were used for reference and validation of the HPRD phosphorylation data. Thus, the entire platelet proteome dataset contained 229 kinases (43.5% of 526 total human kinases), 162 (70.7%) of which have well described substrates in the platelet proteome. Nearly all (216, 94%) have documented phosphorylation sites.

The PlateletWeb knowledge base presents a variety of advanced options for systems biological analysis of platelet signaling. Each protein has been provided with characteristic features from the functional and network context along with the technique for identification in platelets (level of detection). The phosphorylation status of all proteins can be analyzed, distinguishing individually between phosphorylations derived from published literature and those directly measured in human platelets. Furthermore, information on physical properties such as isoform-specific sequence information, presence of transmembrane domains, isoelectric point and molecular weight data are provided and searchable which could be helpful tool for the analysis of Western Blot and 2D Gel experiments. Associations with diseases can be found using a key-word functionality search in the description of proteins. The resource even allows users to combine various search options into a more complex advanced search concentrating on specific platelet proteins.

To better understand the platelet commitment in processes such as immune defense, a functionality search (using GO terms) results in proteins involved in "immune response" and associated functions. Additionally, it is possible to combine GO terminology with the phosphorylation state, level of detection and presence of particular protein domains to retrieve groups of proteins fulfilling the search-defined criteria. Drug information provides additional insight about platelet proteins associated with specific drugs. As an example, results on pharmacological modification by inhibiting prostacyclin receptors retrieve analogues of prostacyclin (Epoprostenol, Iloprost, and Treprostinil). The knowledge base thus allows a comprehensive and detailed analysis of the platelet not only on a single protein level but also on the scale of network regulation and functional association of signaling components. A detailed tutorial on the usage of the platform is presented in Supplementary Material.

The intriguing feature of PlateletWeb, the subnetwork extraction helps to understand the proteins from multiple aspects. For instance, a similar protein can be found performing special functions in multiple networks. The flow of the information in each of the networks could be easily identified using this feature of PlateletWeb. Additionally, the example of LXR provides an insight of how the mechanism can be understood and used.

Motivating the insights on the platelet tyrosine, the conclusion indicates that most of these kinases were found on the proteome level. The detailed analysis of the phylogenetic tree further provided the substantial motivation to this fact. The kinase subtree revealed a very strong representation of Src-family kinases, which have many characterized substrates in platelets and mediate platelet activation through numerous phosphorylation events. In contrast to this, a clear absence of kinase groups such as the neurotropic tyrosine kinases (TRKA, B, C) and neural growth kinases (ROR1) can be observed, as these are kinases with a high specificity for neuronal tissues.

The analysis on the protein domains, especially on the enrichment test suggested that the most often occurring domain is the RAB domain. RAB proteins have a specific function in membrane tethering and fusion by recruiting factors, which interact with SNARE proteins and form SNARE complexes (*Zhang, Naslavsky et al. 2012*). The enrichment of these domains might be explained by their strong involvement in vesicle transport, which is also the underlying mechanism of endocytosis. Additionally, platelets were also found enriched for the endocytosis pathway and the two findings correspond well with each other (*Boyanova, Nilla et al. 2012*).

The SH2 and SH3 domain enrichment is well explained by the fact, that these domains bind to tyrosine phosphorylated residues of platelet proteins after platelet activation transmitted by the kinase Src (*Liu, Jablonowski et al. 2006*).

In another aspect of the analysis on the protein domains, the high interacting proteins with over 50 interactors were considered. Interestingly the placement of the domains was changed, SH2 replaces RAB as the most often occurring domain among highly connected proteins. This change can be explained due to the fact that this domain is an important signal transductor during platelet activation and many proteins after tyrosine phosphorylation can bind to these domains. Therefore, it can be assumed that the enrichment of such domains ensures a more global signal transduction. The analyses underline the tight functional relationship of SH2-domain proteins and tyrosine kinases (*Liu, Jablonowski et al. 2006*) in platelets, which is well reflected by the overrepresentation of tyrosine kinase substrates and SH2 domains in the enrichment analyses.

3.2 Motif analysis

3.2.1 Analysis

To get more insight into the way how information is processed in the anucleate cell type, network motifs which proteins participate in certain complexes of interactions were considered. In order to eliminate the bias of already known biological knowledge when extracting subnetworks, the network motifs were introduced. Although having the knowledge is useful, an unbiased method to search for structural topological elements in the configuration of the network can provide with advanced insights on the complete network. Network motifs can be defined as the unique patterns of interactions between proteins that appear significantly more often in the real network compared with randomized networks. These patterns are considered to be selected by evolution. Therefore, motif analysis introduces new insights on the types of network regulation. In the transcription networks, they were shown to have important information processing functions (*Milo, Shen-Orr et al. 2002; Shen-Orr, Milo et al. 2002; Alon 2003; Mangan and Alon 2003*). This implies that the found motifs might be of importance for the regulation of the observed system. Here, the integrated platelet network consisting of protein-protein interactions (undirected edges) and kinase-substrate interactions (directed edges; arrows from the kinase to the substrate) focusing on a set of 14 motifs were analyzed, which are previously been described to have an impact on information processing in transcriptional networks (*Milo, Shen-Orr et al. 2002*) as well as in signal transduction networks (*Zaidel-Bar, Itzkovitz et al. 2007*).

There are multiple tools available to detect the network motifs – Pajek (*Batagelj and Mrvar 1998; Batagelj and Mrvar 2003; De Nooy, Mrvar et al. 2005*) MAVisto (*Schreiber and Schwobbermeyer 2005; Schwobbermeyer and Wunschiers 2012*), FANMOD (*Wernicke and Rasche 2006*), MFinder software (*Kashtan, Itzkovitz et al. 2004*) to name a few. Alongside, there are multiple plugins available for Cytoscape – Netmatch (*Ferro, Giugno et al. 2007*), CyClus3D (*Audenaert, Van Parys et al. 2011*) are frequently used.

Here, the MFinder software (*Kashtan, Itzkovitz et al. 2004*) was used in order to detect network motifs in the platelet proteome. The software enumerates all n-node patters and classifies them into one of several topologically distinct sub-graphs. The original network is randomized 1000 times keeping the incoming and outgoing edges at each node the same for all the randomized networks. This rewiring of the edges would create complete random networks with random connections for all the 100 randomized networks. Subgraphs that are observed significantly more frequently in the real network than in randomized networks are regarded as network motifs. In the table below, the higher the z-score, the higher the significance of the particular motif in the real network. Scores are shown for all 14 motifs from Zaidel-Bar et al (*Zaidel-Bar, Itzkovitz et al. 2007*). The motif ID indicates the corresponding number of the same motif defined Milo et al (*Milo, Shen-Orr et al. 2002*).

Table 14 – Motif analysis – the significant motifs and their Z-scores

Motif number	***Motif ID***	***N-Real***	***N-Rand stats***	***N-Real Z-Score***
1	110	831	261.9+-18.3	31.16
2	108	128	86.6+-10.9	3.79
3	102	46	15.3+-4.2	7.31
4	238	3432	806.9+-35.6	73.72
5	350	1699	406.9+-130.9	9.87
6	5086	2611	352.0+-60.5	37.32
7	478	128	30.7+-17.5	5.57
8	6558	106	12.9+-4.2	21.95
9	908	50	26.3+-8.5	2.77
10	6604	56	28.6+-10.3	2.65

Motif number	***Motif ID***	***N-Real***	***N-Rand stats***	***N-Real Z-Score***
11	5022	116	17.5+-6.7	14.68
12	5084	829	165.8+-31.4	21.12
13	4574	420	97.3+-22.8	14.13
14	13262	2038	155.0+-34.1	55.16

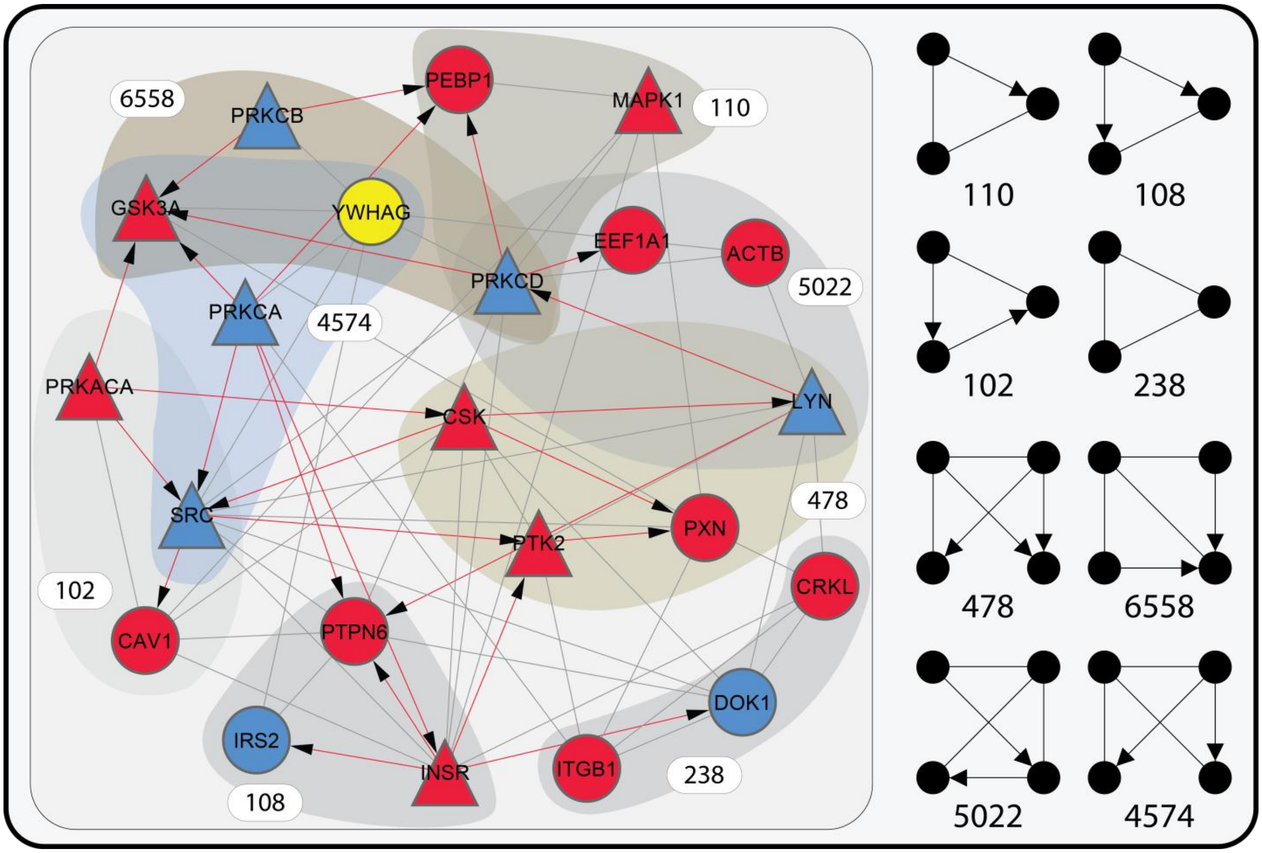


Figure 15 – Motif analysis

The 8 key motifs showcased in a simple module of network

The 8 key motifs with high biological importance and highest z-values are shown in the figure. There are altogether 686 platelet proteins, which build up these motifs. All kinases and proteins shown on figure 2 participate in the most significant eight motifs. The two motifs which occur most often in the platelet network are motif 110 and motif 238, which contains only interactions. The eight motifs allow various kinases to be partners of different logical circuits and allow each kinase to play different roles in different regulation types.

Interestingly, results included two top motifs critical for rapid information processing: One is a scaffold motif which captures one enzyme or kinase by a scaffold protein which then furthermore allows the kinase to choose between two substrates. In Motif 13(4574 in the above depiction) for example one protein (the protein YWHAG from the 14-3-3 family) scaffolds an enzyme and two alternative substrates, in this particular case the kinase are PKC, which alternatively phosphorylates SRC or GSK3A. The main protein could serve as an adaptor molecule for the phosphorylation event, thereby facilitating the kinase activity in the one or other direction. A systematic examination was conducted to see how often a certain protein in the motif appears among all found motifs of this type, and came to a conclusion that in the position of the scaffold protein mainly kinases and actual scaffold proteins were most frequently appearing.

For motif 13(4574 motif) not only the kinases such as FYN and SRC were found but also the adaptor proteins such as GRB2, YWHAG and phosphatases (PTPN11, PTPN1) in the first position in the motif, which plays the role of a scaffold for the enzyme and its substrates. The second position was exclusively taken by kinases, among which SRC, AKT1, INSR and PKC were overrepresented. The third and fourth position was predominantly occupied by the following proteins: CTNNB1 (Catenin), GRB2, CAV1, STAT3. These are the most frequent substrates of kinases from this particular regulatory motif. The same motif, consisting of these as well as other proteins, is found 19 times in the integrin signaling pathway, which indicates the high importance of such a pattern in the interactome.

Motif 110 assembles signaling complexes. For instance, the analysis on ILK PINCH pathway, crucial for integrin signaling(*Lange, Wickström et al. 2009*) (Supplementary Information, fig.S4): the ILK kinase, which phosphorylates GSK3B (also known as GSK3β) is found together with Paxillin – an adaptor protein needed for ILK localization to focal adhesions and downstream integrin signaling(*Nikolopoulos and Turner 2001*).

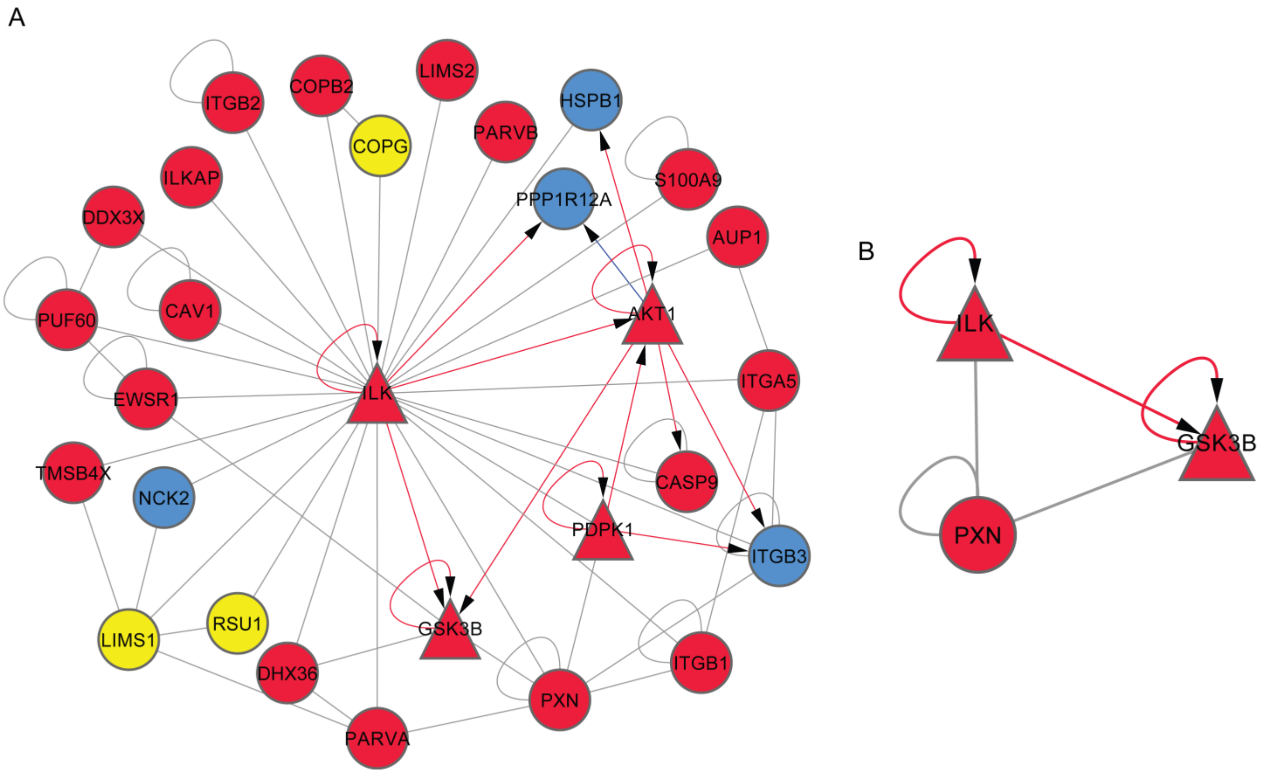


Figure 16 - Motif analysis on the ILK network:

In the above figure (A), all the proteins interacting with ILK (Integrin-linked Kinase), were extracted and visualized according to their interactions and phosphorylation state (red line, phosphorylation; grey line, interaction; blue line, kinase-substrate prediction). The proteins were colored as follows – red, protein phosphorylated in human cells; blue, protein phosphorylated in platelets; yellow, a platelet protein. The kinases are represented by triangles. The motif 110 was found once in the above network (B), containing the integrin-linked kinase (ILK), its substrate GSK3B and the adaptor molecule Paxillin (PXN), which regulates ILK localization to focal adhesions.

Another biologically motif of high relevance is the motif 8 (6558), where an adaptor protein acts as a scaffold for two kinases and their common substrate. Systematical analysis of the proteins contained in the motif, showed that on position 1 there was an actual adaptor protein in most cases and positions 2 and 3 were taken by kinases. A relevant biological example is the binding of DOK2 to HCK and SRC kinases, which both phosphorylate RASA1. Thus, it can be concluded that this motif might be important in regulating platelet cellular processes.

The proteins that were found in most motif structures were taken into consideration and then constructed an interaction and phosphorylation network using these proteins. They build a dense interaction network, and some of the most prominent platelet kinases can be found, such as PKC and LYN.

Motif analysis has further increased our knowledge of what possible mechanisms might be involved in the platelet phosphoproteome and interactome network, pinpointing motif 110 as the pattern with highest significance. Therefore, it can be suggested that motif 110 plays a leading role in cellular signaling of the anucleate platelet. The bingo analysis was performed on all the motifs and the network motif results for 110, 4574 and 6558 are shown in the supplementary material.

To perform the motif analysis a complete interactome network has to be created. This means that if the network is changed by adding or removing the amount of information in it, there is a possibility of the network information to be changed. However, this does not mean the information the motif analysis is providing is falsified. It can be understood as, with the amount of information content we have at hand, the results offer the first hand insight of the network patterns. When further information content is added or deleted with respect to the interactome network, it is then possible to make a comparative analysis and focus on what patterns have actually changed and the implication of the information. Also the comparison analysis on the other species networks on the similar proteins would also yield yet more useful information which might not be easily understood. Additionally, the statistical analysis that is performed decreases any false positives that might be already creeping in to the network. The motif analysis can be thus considered a highly reliable way to identify the network patterns that might be existing in the interactome networks.

3.3 Identifying functional modules using semantic similarity and Heinz (Qualitative analysis)

3.3.1 Qualitative analysis

The extraction of functional modules is a key in understanding the proteins in the context of their interactions. Lots of research was conducted on this fact, however each of the methods that were described have both their advantages and disadvantages. The key for this research is to minimize the disadvantages and bring out the functional modules which provide with most concrete information.

The main problem is that the large proteomic datasets (with over 100 proteins) yield huge networks, which are difficult to interpret and must be considered to further divide them into smaller networks for understanding it. However, in many cases, this results in losing of the important information. Additionally, the proteins are rarely grouped according to their functional relevance. In most of the cases, the proteins bearing similar functionality tend to interact together more tightly rather than the other way round. When the subnetwork is built, the proteins with higher functional relevance are expected to come together and provide the functional module with higher information content in it.

Additional to this, in the proteomic analysis, it’s not always necessary to get the differential expression numbers. In other words, a number defining the proteins role in the network is missed. So, all the proteins should be equally rated to get the interacting network out of all the proteins. However, this does not satisfy the needs of the end users. The difficulty in quantifying the proteins which are identified in the proteomic analysis and extracting the most biologically informative subnetwork prompted to further investigate and improvise the concept of extracting the subnetworks from the proteomic analysis.

The main objective of the qualitative analyses is to derive the biologically interesting subnetworks of interpretable size or of fixed size from large scale protein protein interaction data by logically quantifying the proteins and their interactions. This problem was previously addressed in multiple ways (*Ideker, Thorsson et al. 2001; Dittrich, Klau et al. 2008*). For the first time, Idekar et al has analyzed and expressed this as the problem of finding optimal-scoring subgraph. This problem was then transformed by dittrich et. al, to the well-known PCST problem from Operations Research. With this an alternative NP-completeness proof they could solve large instances of this problem in reasonable computation time. This solution which is retrieved can be considered as the optimal solutions. The algorithm takes another step, and it calculates the suboptimal solutions with given Hamming distances to previously found solution. However this algorithm is given on the basis of the proteins which have the differential expression data associated with it. In our scenario, we do not have the required differential expression data, but just the proteins which are identified in the sample proteins.

In order to address this issue of handing of the proteins, a scoring on the interactions (edges) was made. An edge score can then be used to focus on the most relevant module. The edge score, in this case, should reflect some kind of functional relevance with in the protein-protein interactions (PPI). GO similarity has been developed as measurement of similarity of two proteins in terms of their functional annotation. The main reason behind using the semantic similarity is due to the fact that the proteins with a high functional similarity will be involved in similar cellular processes, thus the interaction between then can be assumed to be more relevant.

Here the Gene Ontology semantic similarity was used as a basis for functional scoring of the PPI. Various measurements of GO similarity have been described in literature, here a score based on the similarity measure proposed by Schlicker was developed.

3.3.2 The Beta Uniform Mixture Model

The Beta Uniform Mixture model (BUM) was introduced by Pounds and Morris et al (*Pounds and Morris 2003*) in 2003.

This approach can be used in microarray or proteomics data where a statistical test is calculated for each gene/protein in the dataset resulting in a single p-value. This p-value is interpretable when using only one statistical test, but in the case of multiple testing the p-values lose their meaning as Type I error rates.

The principle of the BUM is that under a null hypothesis where none of the genes/proteins is of any importance, the p-values will be uniformly distributed. In the case where there are genes with a specific importance to the experiment (holding a signal), there would be an overrepresentation of small p-values and the histogram will represent a peak near the zero. This alternative hypothesis can be modeled by a beta distribution where the p-values are visualized as a mixture distribution of signal and noise component.

3.3.3 Example: why edge scores

A small example here illustrates the importance of using the semantic similarity of the genes and scoring it onto the edges. In this example, let A and B are two proteins which are of interest and are identified in the sample proteins. Both A and B are connected over three different proteins, C, D and E. Considering the edges has no scores on them, the solutions are either ACB, ADB or AEB. However, when adding the biological insights into this small network by using the semantic similarity between the proteins and giving the edges a score depending upon the calculated similarity, the maximum scoring graph will always be the one with the higher connectivity. In our example, the solution would be only ACB as protein C being semantically similar with both proteins A and B.

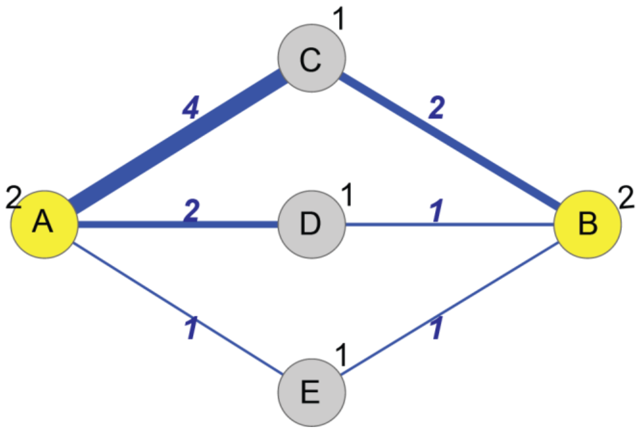


Figure 17 – The schematic showing the difference in the network when the edge scores are included

3.3.4 Calculation of Edge scores

The edge scores of the network are calculated in multiple steps. The complete scenario is presented below in the form of steps, for easier readability.

Step 1: The interactome (original) is randomized twice (random1 and random2) as the key idea is to consider the probability of the observed score showing up in the randomized distribution

Step 2: The semantic similarity is calculated using GOSim on all three interactomes (original, random1 and random2) by setting the ontology to biological process

Step 3: The empirical cumulative distribution function (ecdf) is computed, which when used as function with an object, returns the percentiles. Here, the object is the Semantic similarity scores calculated for the interactome.

Using the statistical software R, the p-values can be achieved by the following:

BiologicalProcessFunctionRandom1 <- ecdf(BiologicalProcessRandom1GOScore)

BiologicalProcessFunctionRandom2 <- ecdf(BiologicalProcessRandom2GOScore)

where "*BiologicalProcessFunctionRandom1*" and "*BiologicalProcessFunctionRandom2*" are the functions that are created for the randomized interactomes which contains the GOSim scores of the Randomized networks.

These functions can now provide the percentiles:

bp_r1percentiles <- *BiologicalProcessFunctionRandom1(BiologicalProcessOriginalGOScore)*

bp_r2percentiles <- *BiologicalProcessFunctionRandom2(BiologicalProcessOriginalGOScore)*

bp_r1s1 <- *BiologicalProcessFunctionRandom1(BiologicalProcessRandom2GOScore)*

bp_r2s2 <- *BiologicalProcessFunctionRandom2(BiologicalProcessRandom2GOScore)*

The first two percentiles, "bp_r1percentiles" and "bp_r2percentiles" are calculated for attaining the edge scores. The next two, "bp_r1s1" and "bp_r2s2" are calculated in order to test if the randomized interactome GOSim scores are bringing any additional effect to the original network. These percentiles are transformed into p-values.

```
bp_pvals_r1p <- 1- bp_r1percentiles

bp_pvals_r2p <- 1- bp_r2percentiles

bp_pvals_r1s1 <- 1-bp_r1s1

bp_pvals_r2s2 <- 1-bp_r2s2
```

Step 4: These calculated p-values can be fit to the Beta Uniform Distribution (BUM) model

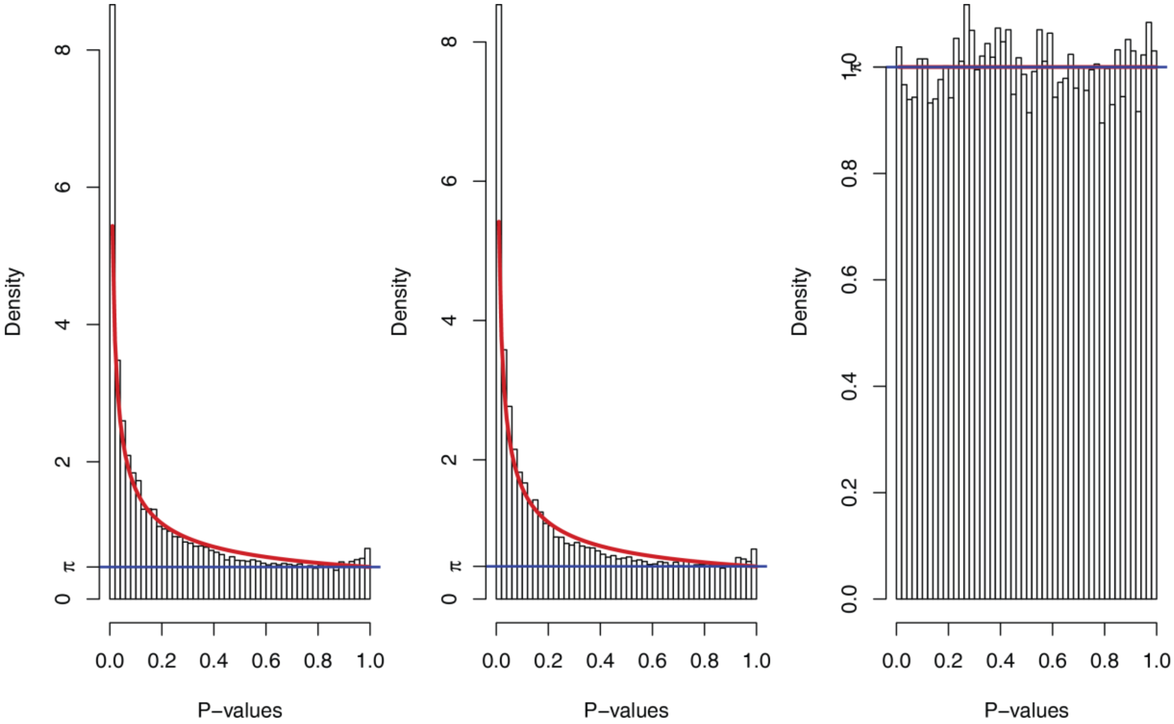


Figure 18 – Beta uniform mixture models and the pi-upper values between the randomized and original networks

The left most figure gives a pi-upper of 0.7 is the model for *bp_pvals_r1p,* the middle one for *bp_pvals_r2p* and the extreme right is for *bp_pvals_r1s1.* The p-values on the extreme right only show the noise and no signal component and the pi-upper is 1.

Step 5: Repeat the steps 2, 3 and 4 by setting the ontologies to Molecular Function and Cellular Component.

Step 6: Aggregate the p-values using first order statistic

Step 7: Fit BUM model to the aggregated p-values to decompose the signal from the noise.

Step 8: Convert the p-values to heinz scores by setting the FDR using the BioNet package.

3.3.5 Example: Conversion of percentiles into p-values

The conversion of percentiles into the p-values requires a little more insight, and this can be explained with a small example.

Initially, using the ecdf function, a cumulative distribution function is computed. In order to calculate the probability of Original interaction GOSim score greater than, say 2.7, it can be mapped onto the curve. Using this value of GOSim as the function parameter, the percentile is calculated. The idea is to see the probability of our observed score of 2.7 showing up on the randomized distribution "P[GOSimscore > 2.7]". It is important to know that the total value of cumulative distribution is always 1. Thus the probability is easily attained by removing the percentile of data from the distribution.

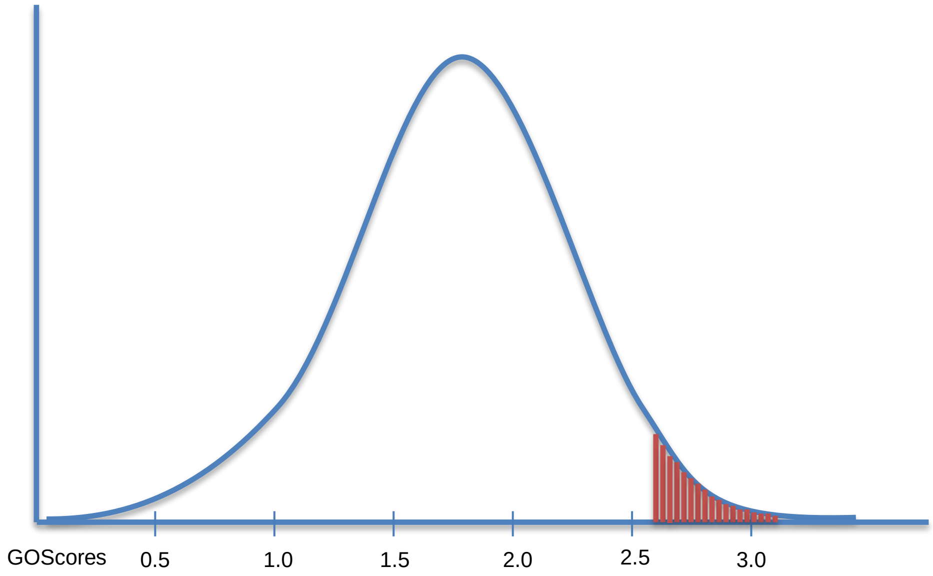
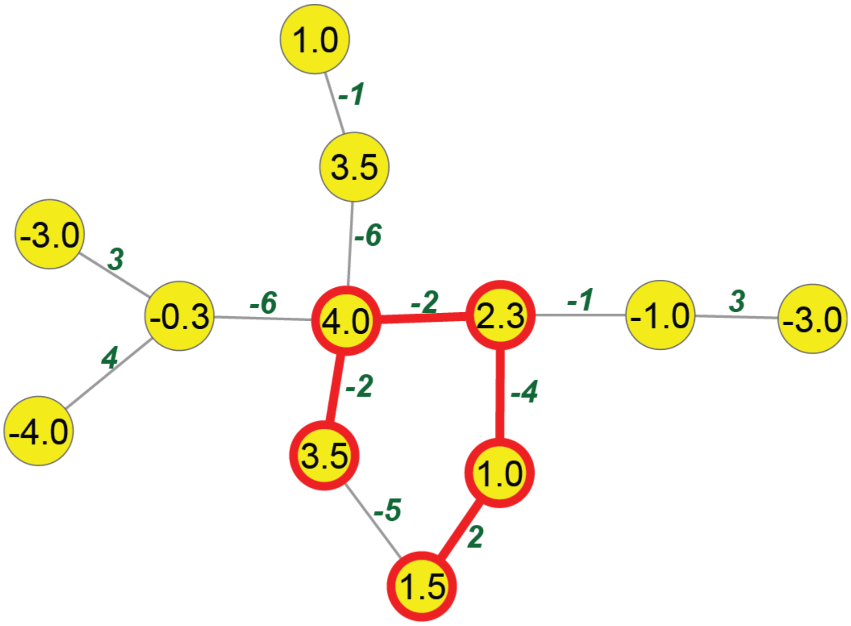


Figure 19 – Figure depicting the way the percentiles are converted into the p-values

3.3.6 Node scores

The algorithm requires the node scores to calculate the maximum scoring subnetwork. The algorithm was previously considered for the data from the microarray experiment or the survival data (*Dittrich, Klau et al. 2008*). In this module, the algorithm is extended in order to identify the functional modules without any known information except the proteins list identified in the proteomic analysis. This implicates a non-availability of the values to consider on the node scores. In order to overcome this issue, a node score algorithm is constructed. The following figure depicts how the node scores are calculated using the already known edge scores which is calculated using the semantic similarity.



$$\Sigma = 3.5 - 2 + 4.0 - 2 + 2.3 - 4 + 1.0 + 2 + 1.5 = \mathbf{6.3}$$

Figure 20 – Calculation of Maximum scoring subnetwork

The proteins identified in the sample are given an adjusted value based on the edge score

Nodescore = -avg(connected edge scores)

This node scores would help to quantify and provide the most optimal values to the proteins identified in the sample. The rest of the nodes in the interactome are given the average of all the edge scores to maintain the background distribution of the network. This would ensure a constant values of edge scores for all the edges and change in the node scores for different samples of data.

3.3.7 Extracting sub networks

The large-scale proteomic datasets generated in recent years have posed a big challenge for developing new strategies to analyze the data in a biologically meaningful way. The large number of proteins identified in samples (>100 proteins identified) make it impossible to isolate functional modules of sizes suitable for analysis. Just by mapping the proteins onto a predefined PPI network is not enough to focus onto the relevantly changed modules contained in the sample. Additional functional information on the interacting proteins is therefore needed. GO similarity has been developed as a measurement of similarity of two proteins in terms of their functional annotation (*Guzzi, Mina et al. 2011*). The rationale behind this scoring system is that proteins with a high functional similarity will be involved in similar cellular processes, thus the interaction between them can be assumed as more biologically relevant. Therefore, one possible solution to this problem is the addition of functional association information onto the edges in the human interactome thereby assigning interaction values according to the functional similarity of the gene ontology associations of the interacting proteins. Thus, the edge score would reflect functional relevance of the particular edge in the PPI.

A score based on the similarity measure proposed by Schlicker (*Schlicker, Domingues et al. 2006*) was developed as detailed in the Experimental Procedures. A functional interaction score was developed based on GO semantic similarity values for all three ontologies (BP, MF and CC). Proteins from a biological sample were assigned values derived from the functional interaction scores, whereas linking proteins from the interactome were given the average of all interaction scores. Both protein and interaction scores were added to the human PPI network, which contains 55,196 edges and 10,688 nodes. An algorithm for detection of functional modules was used for extraction of functionally connected subnetworks, containing the proteins in the sample. The algorithm searches for maximum-scoring subnetworks based on the pre-given edge and node scores. Linking proteins from the human interactome are included in the solution if this increases the maximum score of the extracted subnetwork. These linking proteins might be important players in the overall signaling changes of the analyzed protein sample. Using GO semantic similarity for the functional interaction scores of interacting protein pairs ensures that the algorithm includes paths having high similarity over those having low similarity. Therefore, functionally related clusters of proteins are extracted in the final solution along with linking proteins with a similar function. As proteomic analyses are often fractionated and detection of some proteins proves to be extremely difficult, this approach is useful for unraveling the network context of identified proteins along with proteins missing in the original sample due to technical difficulties. Scores can be adjusted to a threshold, so that more proteins from the original sample are included. Thus, the solution can become more or less stringent, including a higher or a lower number of proteins.

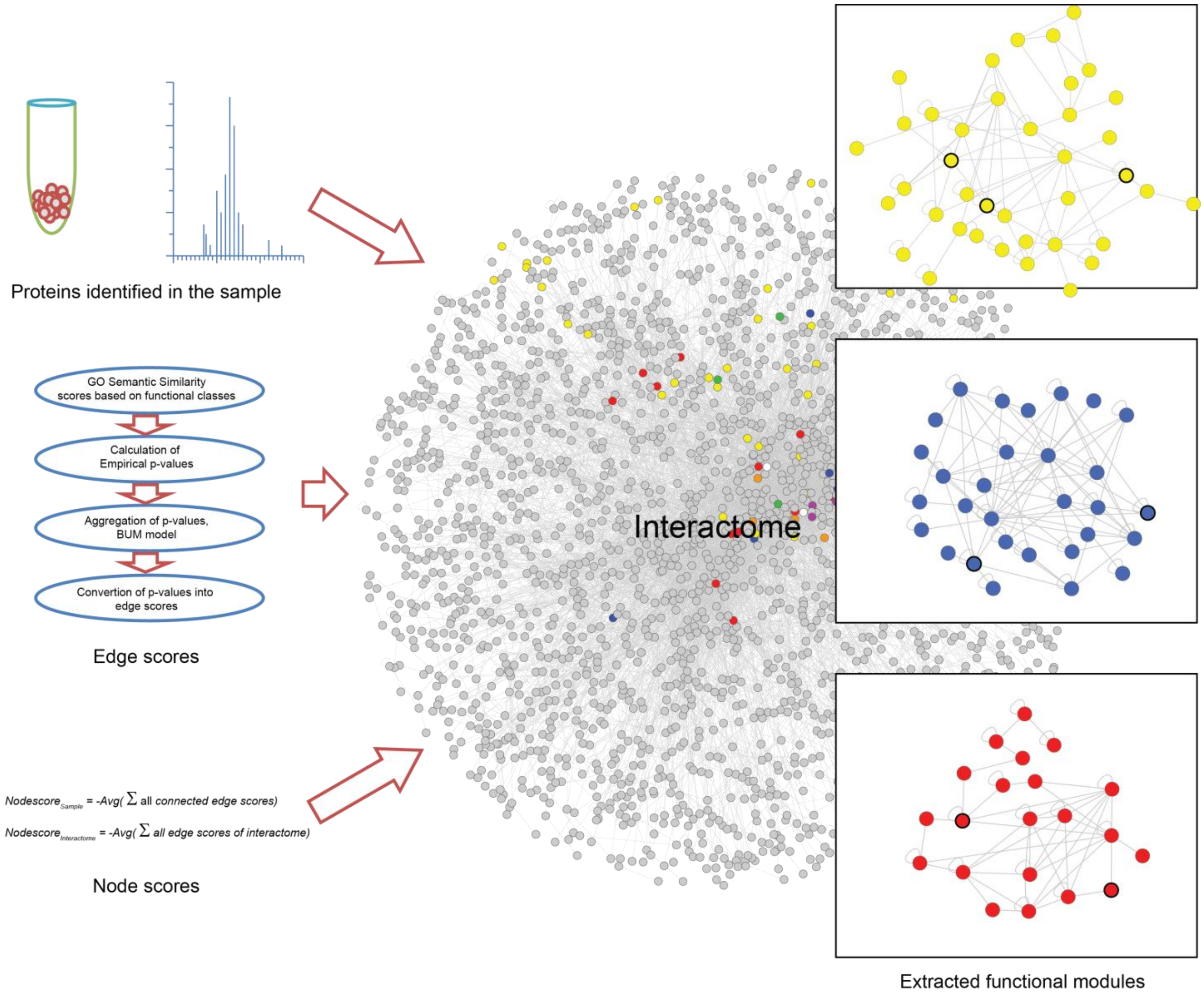

Proteins identified in the sample
GO Semantic Similarity scores based on functional classes
Calculation of Empirical p-values
Aggregation of p-values, BUM model
Convertion of p-values into edge scores
Edge scores
Nodescore$_{Sample}$ = -Avg(Σ all connected edge scores)
Nodescore$_{Interactome}$ = -Avg(Σ all edge scores of interactome)
Node scores
Interactome
Extracted functional modules

Figure 21 – A schematic of the idea for identifying functional modules using edge and node scores

3.3.8 Beta-mixture model (BUM) of Biological Process transformed scores

To combine GO semantic similarity of edges with the topology of the human interactome network a new score monotonously dependent on the original Schlicker scores was introduced. This score allows a natural way of integrating multiple sources of information and is therefore very flexible. A combination of p-values representing different data can be smoothly integrated into the framework. The scores are also adjusted to the functionality of the human interactome. For the calculation, Schlicker BP score was calculated for the human interactome edges and a randomized network of the same size but a different topology (same number of nodes and edges, but rewired). The original network scores and the scores from the randomized one were used for obtaining empirical p-values. A beta mixture model (BUM) was then applied to determine the information content and the signal/noise ratio. The distribution of p-values in this model consists of noise (a uniform distribution (under the blue threshold =Pi-upper)) and signal (beta distribution).

The obtained p-values could be fitted to a beta uniform mixture model (BUM) (Figure 2A) indicating that the human interactome contains interacting proteins with high functional similarity significantly more often than the randomized network model. The *Pi*-upper value (the threshold separating signal from noise) was 0.46710439372661.

In the case of the randomized network, the obtained p-values were uniformly distributed (Figure 2B), which indicates the lack of any relevant functional dependency for nodes from this network. The *Pi*-upper value in this case was 1. The quality of model fit in both cases was assessed by QQ Plots (Supplementary Figure S3).

Scores were further obtained from GO semantic similarity values of MF and CC based on the Schlicker algorithm. Then, a new score was proposed, which smoothly integrates all derived p-values from the three ontologies into a single combined score, further used in all our analyses (see "Experimental Procedures, Calculation of Functional Interaction Scores").

When comparing the information content of each individual score, presented as the 1 - *Pi*-upper value of each BUM model, BP showed the highest signal content, followed closely by the combined functional interaction score based on all three ontology scores (Supplementary Figure S4).

Differences between the three separate ontologies could be analyzed after running the heinz algorithm based on the edge scores only without the node scores and limiting the resulting network to a size of 200 in each case to ensure a reasonable comparison. Thus, there was no need for changing the False Discovery Rate (FDR) parameter as the algorithm searches for the maximum scoring subnetwork limited to 200 nodes. Results indicated a very low overlap between the three ontologies. While BP and CC had 63 proteins in common (%), MF was very distinct from both with only 10 and 13 proteins overlap. This might be due to the nature of the terms in Molecular Function, which focus more on the chemical properties of proteins and isn't necessarily associated with the same biological process or compartment in the cell. Thus, MF represents a different aspect and contains other information not necessarily covered by BP and CC. The distinct nature of all three ontologies made it reasonable to combine them in a more suitable score, which would contain information from all three ontologies. Therefore, aggregating the pvalues, a score which could be used as a combined ready score for the network was obtained. Results show that this score is mainly based on BP as the functions of the two resulting networks are very similar. This might be because BP has the highest information content and therefore influences the final score to a large extent.

Multiple analyses were performed taking different cases into consideration. The heinz algorithm was run considering the size parameter of 200 and giving the nodes a score of 0. The edge scores are given the originally derived scores for each of the ontologies without the change in the FDR. This revealed that the overlap between the modules derived between BP, MF and CC were very low in the edge heinz. There were only 4 proteins which were common to all the three ontology runs, which when checked were belonging to the proteasome complex. When checked individually between the ontologies, a higher overlap was found between BP and CC (63 proteins). Also, MF represents a different aspect and contains further more information which need not have to be necessarily covered in the other two ontologies. When the aggregated p-value of order 1 was taken, it was close to the BP, which can be interpreted as the algorithm is considering the most information from BP and combining it with the CC and MF, as BP is with a very high information content.

In order to validate the algorithm when induced with semantic similarity, vigorous testing and analysis was performed. This also proved the point that this algorithm can be used in the network of any size. The following case studies provide the results when the algorithm is used with different sizes of the network.

3.3.9 Case study 1: Embryonic stem cells for different ontologies

In order to understand how the algorithm reacts when showcased with the same network on three different similarities (BP, MF, CC), the analysis on the study from Brill et al (*Brill,* *Xiong et al. 2009*) on embryonic stem cells and their phosphorylations was conducted. Using the identified proteins as the nodes of interest, the node scores derived from the pre-calculated edge scores were included. The size of the resulting network was once again restricted to 200 nodes. The aim was to see how well the go similarity scores concentrate the network onto biologically relevant information. There was only a very small overlap between results with biological data (ESC data, Figure 22) and results with no biological information (only edge scores, Figure 23) indicating that there is no bias of the edge scores to the network. Furthermore, using the sample proteins gives more specific information than taking only edge scores, as it is focusing the network onto a set of nodes which are of particular interest and their function. The overlap between the separate ontologies was a lot higher in each of the cases pinpointing that node information combined with edge data gives a more detailed and consistent biological results than using only the go similarity scores. The ESC resulting module shows functional clusters, characteristic for embryonic stem cells, such as translation initiation factors, nuclear transport, cell cycle/division and MAP-kinase-kinase-kinases (MAPKKK). These clusters are clearly different from the clusters obtained by using only edge scores without any node information (Figure 23). In this solution there are only complexes with a very high functional similarity.

A set of 200 proteins were taken and the algorithm was implemented with multiple criteria. The following table shows a matrix of different criteria and the overlap of the proteins in each of the cases. Only the edge information of the proteins was taken and the algorithm produced the networks with the following count of proteins in the resulting network.

Edges Only	Aggregate Original interactome 1	Aggregate Original interactome 3	Biological Process	Molecular Function	Cellular Component
Aggregate	200	100	148	25	78

Edges Only	Aggregate Original interactome 1	Aggregate Original interactome 3	Biological Process	Molecular Function	Cellular Component
Original interactome 1					
Aggregate Original interactome 3	100	200	114	16	70
Biological Process	148	114	200	10	63
Molecular Function	25	16	10	200	13
Cellular Component	78	70	63	13	200

The table showcases that the aggregated order 1 is similar to Biological process than the other two (molecular function and cellular component). Also, the molecular function contains the most distinct information with very low overlap among all the three ontologies. The solutions for each of the ontologies are also very different reflecting the variability between information coming from the three ontologies.

The similar analysis was performed with same edge scores, however this time with the nodes in the back ground. The algorithm was made run on multiple criteria again, and this time the results conclude in a different way.

Edges with nodes background	Aggregate Original interactome 1	Aggregate Original interactome 3	Biological Process	Molecular Function	Cellular Component
Aggregate Original interactome 1	200	26	88	-	30
Aggregate Original interactome 3	26	200	51	-	52
Biological Process	88	51	200	-	-
Molecular Function	-	-	-	-	-
Cellular Component	30	52	-	-	200

The number of proteins was once again, 200 and this was once again compared in the matrix format. In this scenario, the number of proteins in the solution modules was too little in most of the cases. Also in the cases like molecular function, the algorithm could not give any module out and the algorithm continues to run to find out the functional modules. The algorithm was made stopped in the middle for the molecular function as it is clear that there are no or none of the functional modules existing within the set of proteins on the molecular functional level. Additionally, the relation between Aggregate interactome 1 and 3 provides a drastic change. The following figures show the gained results from the algorithm runs.

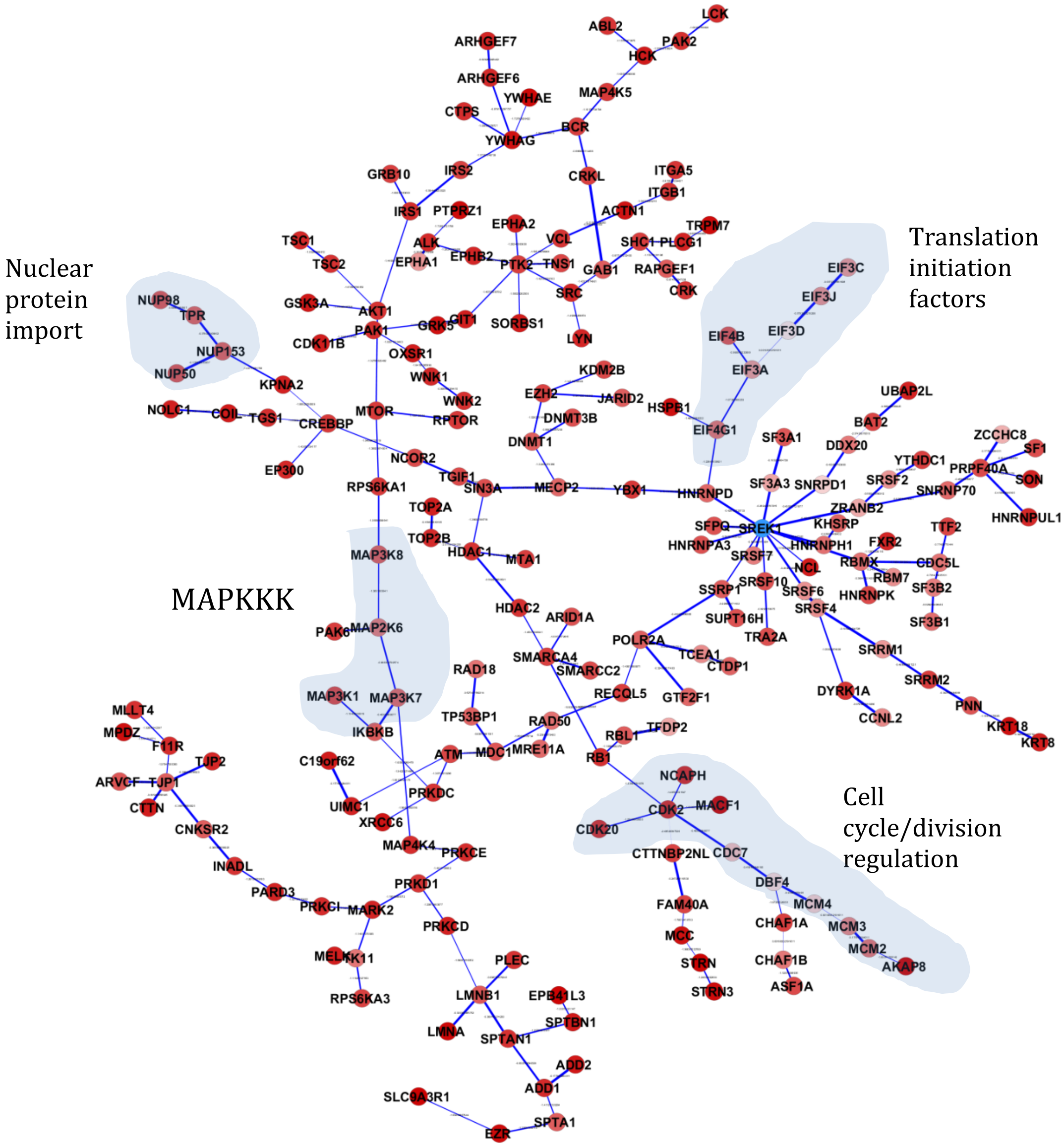


Figure 22 - A network solution of ESC proteins with edge scores

This network depicts the solution achieved from the ESC proteins with the edge scores. The functional interaction scores are equal to the aggregated p-values. The size of proteins is 200 and the subnet edge score was -192.542332563.

The red shades represent low to high scores from light red to dark red. Edge scores are also used. Linking proteins from the human interactome are shown in blue.

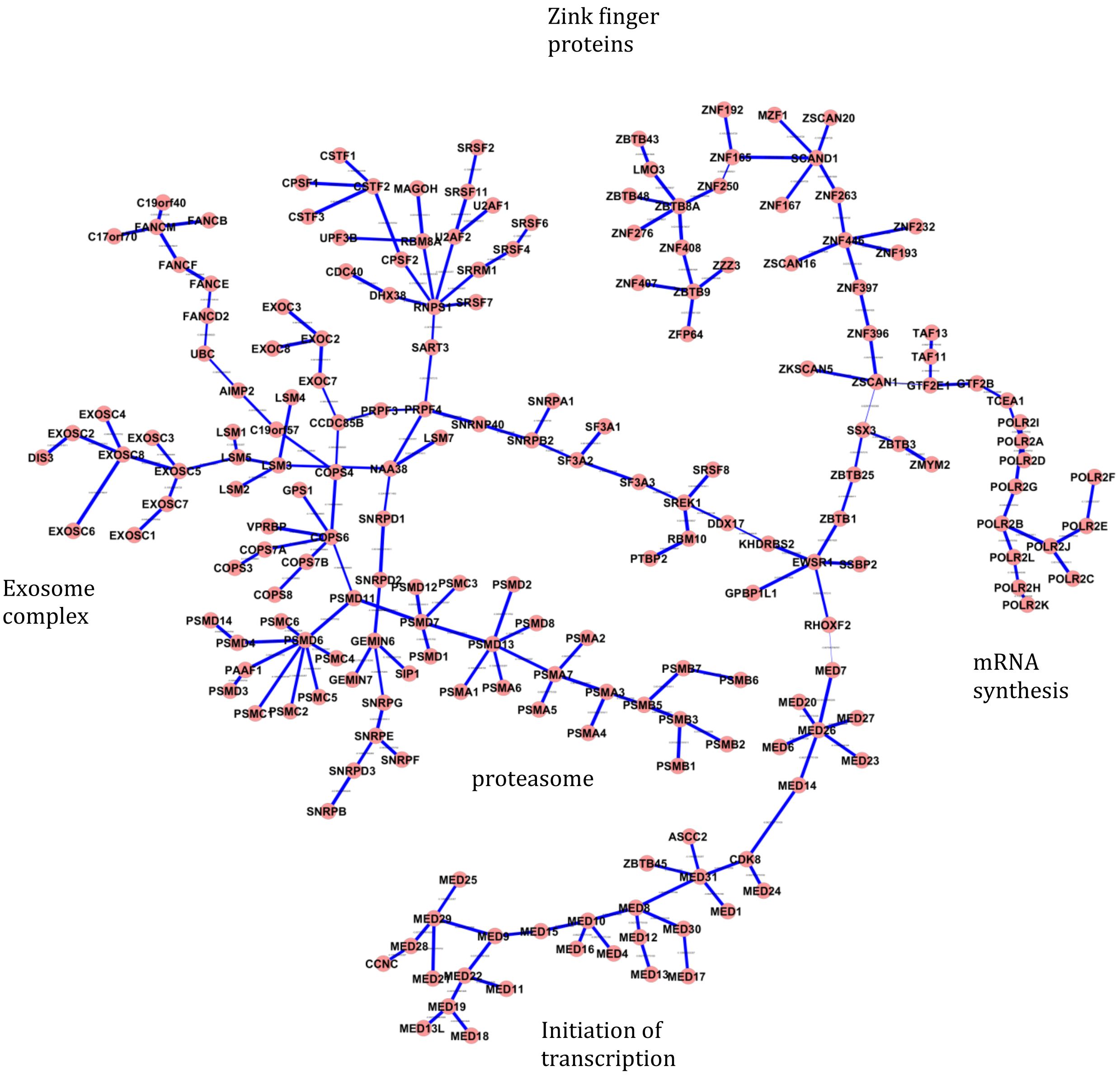


Figure 23 - A network solution using only edge scores

The figure shows the network solution only using the edge scores. Again, the functional interaction scores are equal to the aggregate p-values. In this scenario only the edge scores are used without the node scores. The subnet edge score total was -21.2314232112.

3.3.10 Case study 2: Extraction of network modules using qualitative proteome data

The main advantage of heinz algorithm is that it can be easily possible to visualize the proteins identified in the proteomic analysis even though they are not directly interacting (connected) with each other. The heinz algorithm together with the functional similarity between the proteins helps to connect these proteins in more optimal way, keeping the functional integrity of the analysis intact. In order to validate this, the function was applied on two smaller networks (*Yang, Xiao et al. 2006*), (*Liu, Song et al. 2008*).

In the first case, a human gastric epithelial cell line (AGC) was analyzed after infection with avian influenza virus (H9N2) using mass spectrometry identifying 22 proteins. These proteins were used as the set of interest and analyzed with and without edge scores. In the case without edge scores, there was no score applied to the edges and the nodes that are identified in the analysis were given a constant value of +5 to assure that they all appear in the final network solution. The remaining nodes in the interactome were given a value of -1 (a negative value, indicating that the proteins might or might not be actively participating in the network). The resulting network contains more basic information driven only by the nodes from the sample and not integrating any additional biological data. In the alternative case, the aggregated pvalues were used to combine the three ontologies and used as the edge scores, and the nodes of interest are then derived from the edge scores. Additionally, a constant of 10 was added to these node scores to ensure that all proteins are in the resulting network. The result with edge scores provided more defined pathway connecting keratins to the rest of the proteins (over RAF1).

Proteins obtained with proteomics analysis can be visualized in a network context, even though they are not directly connected with each other. Introducing functional interaction scores to the module detection algorithm helps to connect these proteins in an optimal way based on the functional context of the protein modules and the scores of the interactions connecting them.

In order to test this, we applied the algorithm on a small biological sample from a human gastric cell line (AGC, originating from a gastric epithelial cell line) analyzed after infection with avian influenza virus (H9N2) using mass spectrometry. The study identified 22 proteins (*Wu, Noh et al. 1996*). Two separate analyses of this protein sample were performed to investigate the advantages of functional interactions scores: including and not including interaction scores in the module extracting algorithm (see "Experimental Procedures"). When adding functional interaction scores, we used the previously described method for combining the three GO similarity scores and proteins from the sample assigned a node score derived from the edge score (Figure1B). High constraints for the node scores were used to ensure that all identified proteins are found in the resulting network (see "Experimental Procedures, Constraints of the algorithm solution").

The method with functional interaction scores (Figure 3B) yielded 39 proteins, whereas the method without interaction scores extracted 37 proteins (Figure 3A). Both networks contain linking proteins and paths, which partially overlap (14 paths and 5 linking proteins are same). There are 28 proteins in common for the two solutions but there are also unique proteins, which appear only when functional scores are introduced to the edges (AKT1, EPB41, GNG4, HRAS, HSPA4, JAK2, KRT5, PDK1, RAF1, YWHAB). The number of literature citations for each interaction is very low in the network without functional scores. In the network with scores there are five interactions with more than one citation, indicating the algorithm chooses well-annotated interactions.

3.3.11 Case Study 3: Human Blood cell constituents

To test whether our method is applicable for differentiation of various cell types in terms of functional modules, we used a proteomics study by Haudek et al (*Haudek, Slany et al. 2009*), where mass spectrometry of human blood cell constituents was performed.

The module detection algorithm was performed on the protein samples of all 7 cell constituents separately: plasma, T-cells, neutrophils, monocytes, platelets, erythrocytes and PBMC (peripheral blood mononuclear cell). In the next step, the resulting networks from all blood cell types were analyzed for enrichment of biological processes. GO terms containing less than 3 or more than 600 proteins were excluded from the analysis. The significance threshold of terms in this case was set to a corrected p-value of least 1 × 10-5 in at least one of the cell constituents. Clustering of GO terms was performed with hierarchical clustering based on Euclidean distance calculation.

Blood constituents clustered in a way consistent with their developmental stages. As expected, plasma proteins build an out-group, distant from all other blood constituents as they are not a cell type. Two big groups are formed (platelets, neutrophils and erythrocytes vs. monocytes, tcells and PBMC), which are consistent with hematopoiesis stages.

There are terms specific to platelets like the term “platelet activation” and hemostatic processes, while other terms such as “cell killing” and “leukocyte-mediated cytotoxicity” are enriched in monocytes, which biologically differentiate these from other blood cell types. Ubiquitin and proteasome processes are highly enriched in three cell types: T-cells, monocytes and PBMC (which are a mixture of cells with a round nucleus, including T-cells and monocytes). The term “translation” is overrepresented in T-cells and PBMC when compared to all other constituents. The biological process “generation of precursor metabolites and energy” is overrepresented in monocytes.

In a second analysis we included significantly enriched terms in exclusively one cell type with a p-value of at least 0.05. Specific functions for each cell type could be extracted, indicating the resulting modules are cell-type specific and focus on the characteristic functional modules of the given cell type (Supplementary Table S1).

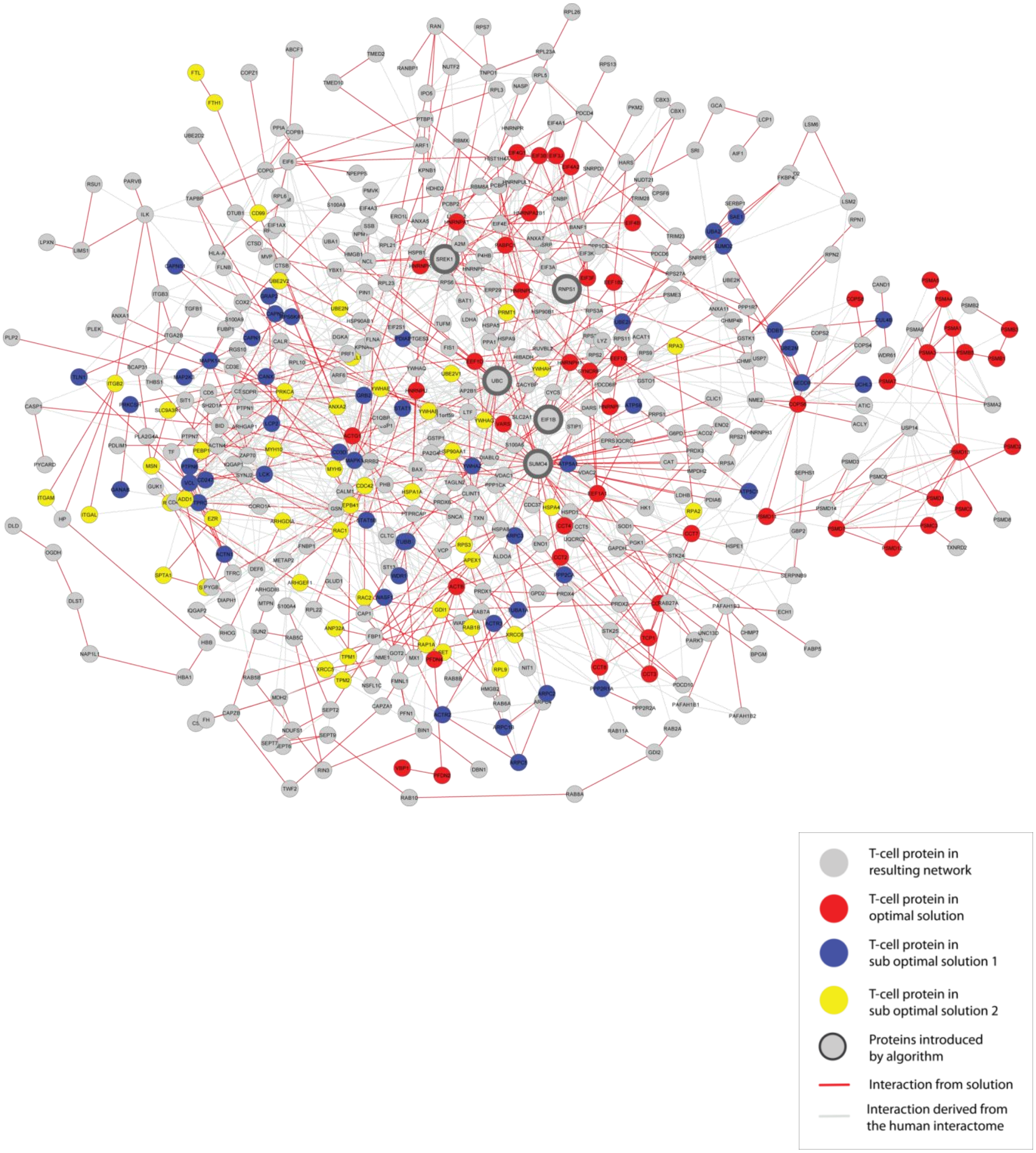

T-cell protein in resulting network
T-cell protein in optimal solution
T-cell protein in sub optimal solution 1
T-cell protein in sub optimal solution 2
Proteins introduced by algorithm
Interaction from solution
Interaction derived from the human interactome

Figure 24 – Human blood cells constitutents; the protein network identified using the sample of proteins

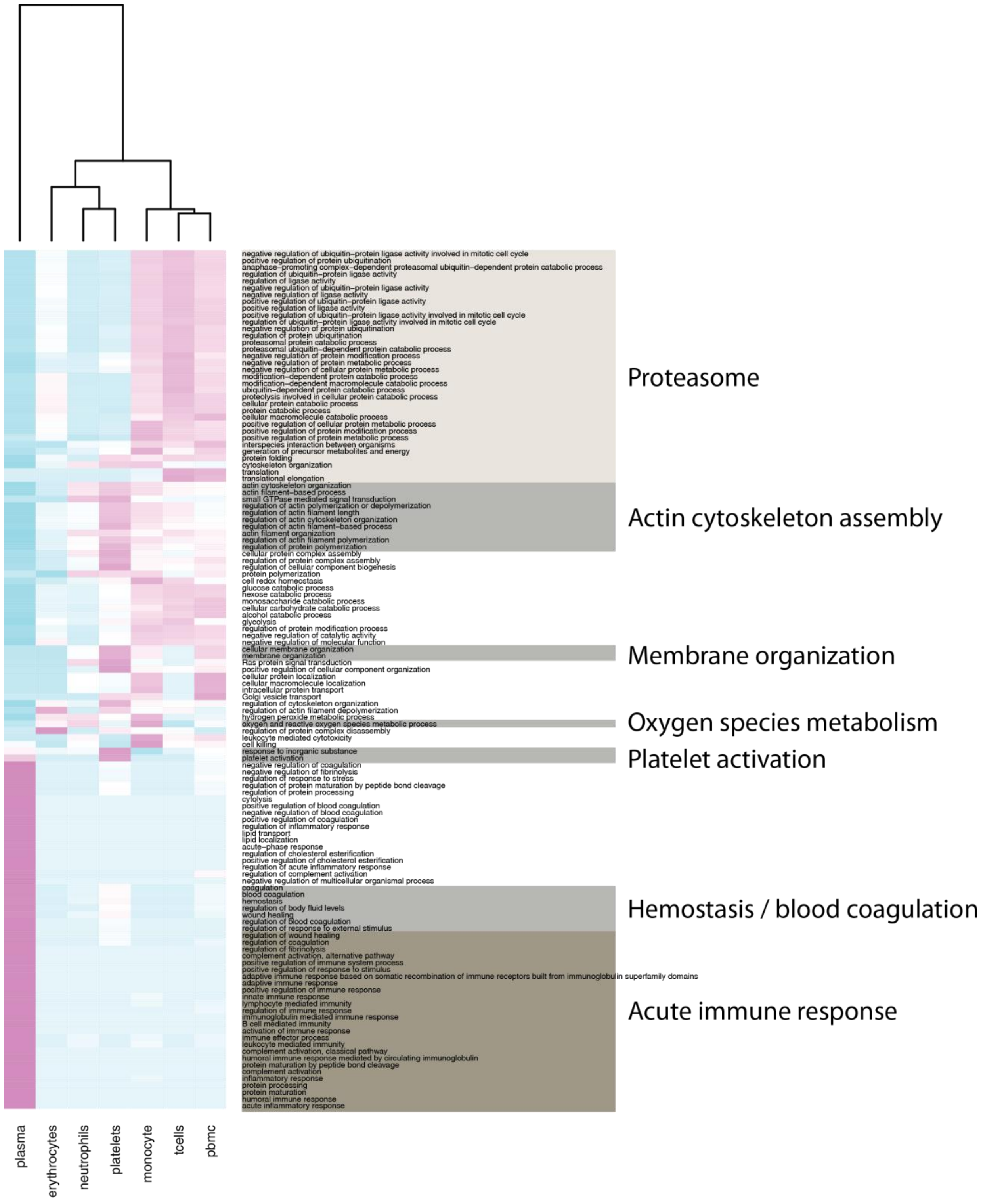


Figure 25 – Heat map of the human blood cell constituents; Testing of the module detection method for various cell types

3.3.12 Analysis on the concept of functional modules

The concept of extraction of the proteins of interest, and trying to get to see the functionality they possess is a vast study and loads of new ideas break open to investigate this issue. The interconnected proteins may or may not follow a pattern which makes the analysis difficult to interpret in first place. It's highly impossible to formulate and interpret the extraction and visualization of the proteins as the needs of the end users differs at different points of the experimental analysis. However, this limitation does not stop from investigating new ideas on extracting the functional modules using statistical and mathematical functions, which might be repeatedly used for any scenario.

An extreme research is being done in trying to extract the functional modules however every method that come across has both the advantages and disadvantages in using it. In this research the proteins were scored as a function of the interaction score. The main power of having this formulation is that it provides the ease of using this on any protein sample as it is pre calculated. Additionally, the nodes with high scoring interactions are downgraded so that they are not dominating the overall possible solution. As an example, the proteins with higher number of interactions might come up more often than expected in any protein sample taken. This might bias the results of the functional module. Downgrading these hubs according to the network structure can provide the actual result which might be on higher biological relevance.

After analyzing the performance of the method on small networks, our next step included large-scale proteomic studies of blood constituents. The overexpression of the term "translation" in T-cells might be explained with the fact, that there is a high protein turnover and proliferative events in these cells causing an increased protein production. Another interesting aspect was the enrichment of proteasome and ubiquitin processes in particularly three cell types: T-cells, monocytes and PBMC. The existence and role of the proteasome in immune as well as non- immune cells has been extensively reviewed by Ebstein et al (*Ebstein, Kloetzel et al. 2012*). Interestingly, these three cell types constitutively express three proteasome subunits (PSMB8, PSM10 and PSMB9), which are normally not included in the proteasome but induced after stimulation. Thus, they build the so called immunoprotesome, which plays an important role in antigen presentation by MHC class I molecules, cell proliferation, cell signaling and cytokine production (*Ebstein, Kloetzel et al. 2012*). This type of proteasome has been identified mainly in dendritic cells and antigen presenting B-cells. As monocytes replenish the pool of macrophages and dendritic cells in the body, the immunoproteasome is found in these cells as well. T-cells show a constitutive expression of all three molecules and the immunoproteasome facilitates protein homeostasis and cell proliferation in these cells (*Zaiss, de Graaf et al. 2008*). PBMCs are a mixture of macrophages, monocytes and lymphocytes therefore it is logical that they would also have this function. All three proteins are present only in the described three cell types, and are lacking or not fully expressed in the other blood cells. This might explain why there is an overrepresentation of proteasome processes when compared to the profile of other blood constituents.

3.4 Phosphoproteomics (Quantitative analysis)

3.4.1 Workflow

The understanding of the proteins and their functional involvement in the network requires in depth research and thus provokes to invent new methods. Additionally, the kinase substrate relationships in the network context provide with a vast knowledge on the signaling networks in the context of the phosphoproteomics and its analysis. The heinz algorithm discussed in the previous chapters would help in determining the underlying network signaling, however, this does not necessarily provide the biologically significant information. The quantitative analysis performed in mass spectrometry provides the new information about the proteins and the differentially expressed quantified value.

This quantitative information gained from the analysis can be used as the node scores for the heinz algorithm. Additionally, site specific information is induced and added into the node scores provide with highly significant information as the resulting modules. In order to test this concept and idea, a network which has all the phosphorylation sites from the dataset that are associated with kinases were taken into consideration(*Rigbolt, Prokhorova et al. 2011*). Additionally, the kinase information is added from the PlateletWeb database in order to provide the real kinase substrate relationships and see the real signal flow. This would then become the additional information provided to the data set, which is not either detected or specified in the proteomic analysis performed on the proteins.

The methodology of quantitative analysis of the dataset using heinz consists of various steps, which are detailed in the steps below.

The analysis was performed on the research resource (*Rigbolt, Prokhorova et al. 2011*), and the genes were then mapped to the PlateletWeb dataset. This was important because the dataset contained the mappings of the IPI, however, the PlateletWeb was using the Gene identifiers in order to identify the gene uniquely. The dataset defines profiling of the proteins in two different conditions (treatments), NCM and PMA. For this analysis, the NCM was used as it contains the information and was well defined in the dataset. The dataset provided with 4 different values at 4 different time points defined for NCM which was taken into consideration.

Furthermore the dataset defines class1 sites, which were also used to filter the dataset as it provides with most concrete dataset for the analysis. This dataset once mapped with the gene identifiers can be ready for the step 2 of the analysis.

All the mapped proteins which are under NCM time line and have the class1 sites can now be made ready to convert into heinz scores to get the functional modules out. For this an additional 1 is added to the complete dataset so that the logs and sumlog of the data can be calculated. This marks as an important step for this method as the sumlog of the differential dataset of 0 is not defined. In other words, if there is no change recorded in the dataset, the log value of this number tunes to undefined value, which is not an intended scenario. In overcome this fact, an additional value of 1 is added to the complete dataset, which would mark the same factor without a huge difference in the logarithmic values. Later, just to figure out the regulation, the absolute log value for all the 4 time lines are taken. This would mean that there are no negative values for the down regulation of the data.

The site information has to be added to the proteins that are identified in the dataset. At first the discrimination between the sites having a mapped kinase from the interactome, and the sites not having a known kinase has to be made. The emphasis is given to the sites with associated kinases. This will help to have only the sites with the kinases in the network (this will be the edge file for heinz). Additionally, it's important to note that there are not edge scores given, so the edge score is considered to be 0. The largest connected component of the network in Cytoscape which is containing the interactome kinases and the dataset kinases and proteins is used as the edge file for heinz.

Each of the proteins might have multiple sites, and here only the maximum scoring sites was considered when more than 1 site is given for a protein. To achieve this, the sumlog is calculated for all the 4 values and the maximum sumlog between two sites was taken into consideration. This can be also termed as "Sumlog of the rows (4 timelines) and then the maximum of the sumlog columns (sites)".

An FDR for each of the 4 absolute logs was provided individually so that when the subnetwork is extracted by the algorithm, the number of positive nodes is at least 50 in each case. This file can be termed as the input node file for the heinz algorithm.

Finally, with both the node file and edge file (even though its 0 in this case) intact, the heinz algorithm can be executed in order to find the maximum connecting subnetwork on all the 4 different time lines. With the resulting information, the scores for all the 4 different points are then extracted and combine the resulting protein networks into one single network. This can be visualized using the Cytoscape in all the 5 different time points.

This methodology when performed resulted in the information content, which is described below.

From the dataset, 67 phosphorylated proteins were resulted of which 14 are kinases and 120 phosphorylation sites with kinase information was also obtained. Additionally from the human interactome (PlateletWeb), 33 kinases were added building up 159 phosphorylation events. This numbers might not be a huge exploration on the first sight, however, when all the 4 time lines are visualized; this revealed many patterns in the networks which can be further investigated.

An output figure for the 24 hours timeline is shown below (Figure 23). The colors represent the absolute logarithmic values (log2) of the phosphorylation ratio between the measurements at 24h when compared to the control. The interval of values is depicted in a color scale from blue (low value) to red (high value). Nodes represent the maximally changing phosphorylation site. Kinases from the human phosphoproteome extracted from literature information (PlateletWeb knowledge base) are shown as yellow triangles. The network allows detailed investigation of the signal flow during embryonic stem cell differentiation with special focus on single phosphorylation sites and clusters of phosphorylation sites changing in a similar manner such as the phosphorylation targets of CDK2. Thus, integrated network analysis of the phosphoproteome helps to identify changes not only on a single protein level, but also indicators for network dynamics which could be shown in the example given in Figure 24.

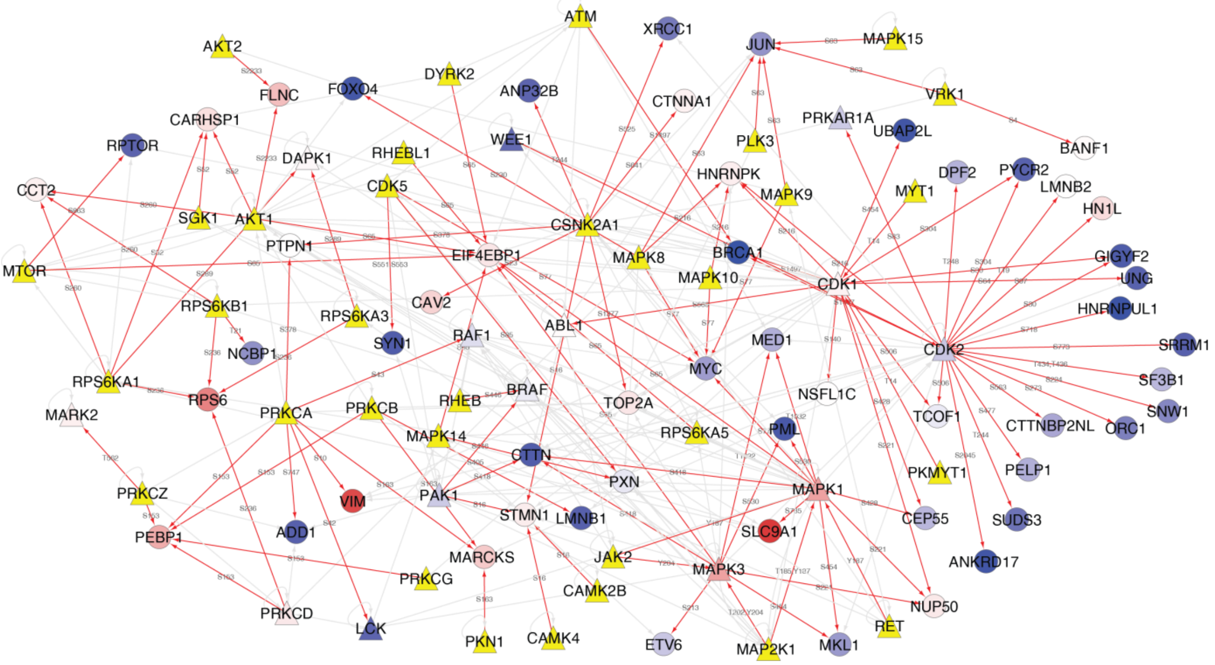


Figure 26 – The module at 24 hours of time line

To further analyze the results, we focused on a classical pathway playing a crucial role in growth and cell division of hESCs. The Raf/MEK/ERK (RAF1/MAP2K1/MAPK1) pathway is visualized with the changes of phosphorylation at each time point. This signaling module is only completed after adding the MAP-kinase-kinase (MAP2K1) and its phosphorylation by RAF1 (*Wu, Noh et al. 1996*), both derived from the human phosphoproteome. Moreover, regulatory switches are revealed by following changes in signaling of a specific module of CDK2 phosphorylating SNW1 at S224 (*Wu, Noh et al. 1996*) (max 1h) and ORC1 at S273 (*Mendez, Zou-Yang et al. 2002*) (max 6h). SNW1 is a coactivator that enhances transcription from some Polymerase II promoters (*Zhang, Dowd et al. 2003*), while ORC1 initiates DNA replication (*Hemerly, Prasanth et al. 2009*). A recent study revealed that phosphorylation of ORC2 Protein Dissociates Origin Recognition Complex from Chromatin and Replication Origins (*Lee, Bang et al. 2012*). This phosphorylation was mainly on threonine sides, but they were regulated by the CDK2 kinase, thus, the phosphorylation here may also play a role in dissociating ORC1 from chromatin.

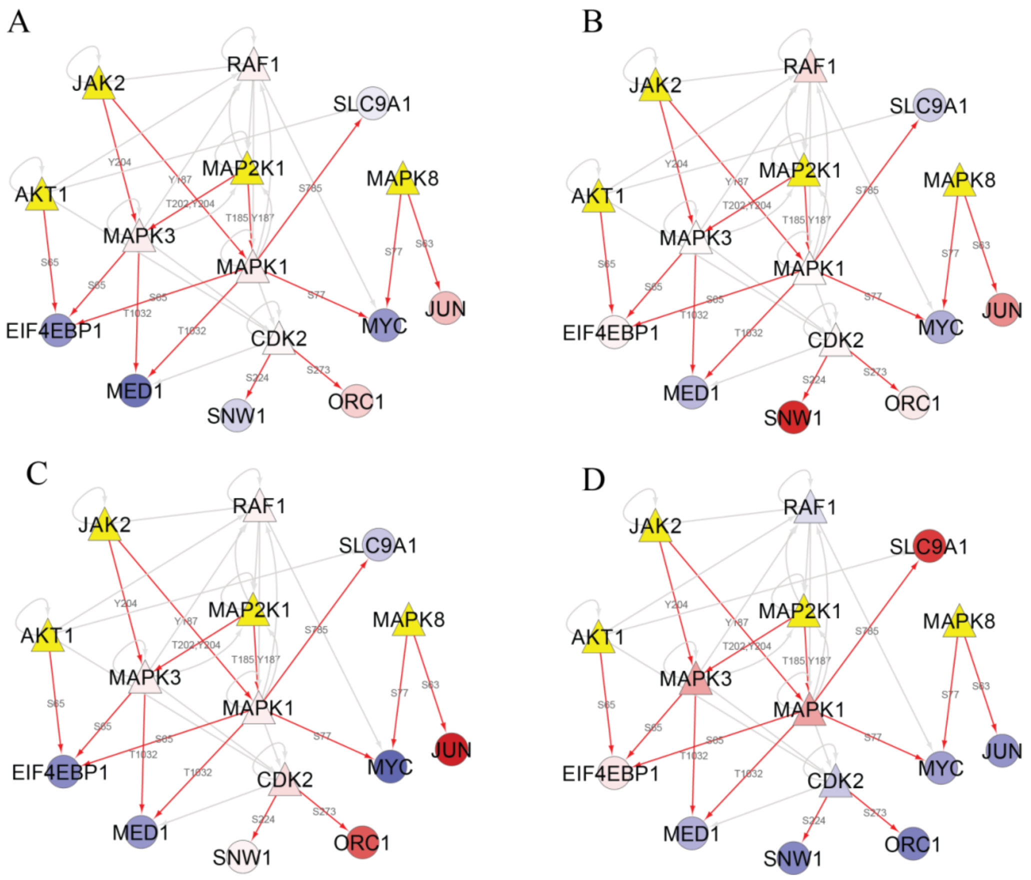


Figure 27 - Four time points of quantitative phosphoproteomics analysis of the RAF/MEK/ERK pathway

The methodology is first of its kind and takes understandably long time. This can, however be fine-tuned to the needs and can also be optimized with respect to the resulting data. Various added values can be achieved using this method. One of the most prominent usages would be to have a detailed analysis of regulation of phosphorylation sites which is understandably lacked in the original dataset. This dataset can be very well visualized in the context of protein-protein interactions. Apart from the site specific information between the proteins, the additional interactions and their associations can be easily identified in the context of the networks. Secondly, the temporal changes in particular sites (time point changes) can be easily identified and investigated further. Furthermore, the phosphorylation events are added so that the regulation in the dataset can be analyzed; for example the kinases and the proteins are available but the link between this may be missing in the dataset. Another key factor which makes this an active methodology is because of the capability to visualize and track the signaling flow inside the pathway. This from a different perspective, the motif analysis is simplified. Additionally, it provides the power to identify individual modules, extract them and further analyze depending upon the context. Apart from all these important usages, it is also possible to identify the clusters of phosphosites which are up or down regulated in a similar pattern.

With all the above discussed usages, the method can be termed as a value added service that can be brought along with the heinz algorithm and PlateletWeb dataset.

4 Discussion

Systems biology focuses mainly on the complex interactions with in the biological systems, which is not possible without proper informatics associated with it. Informatics can be considered as a service, which can be easily induced into any science to help it carry on the research quicker and simpler. In general, the clinic directors hire the doctor students for the tasks associated with the biology. However, the systems biological tasks require an inner depth of knowledge on multiple informatics aspects which requires the informaticians to get in. A cleverly established database which can be easily expandable and scalable with the new information, easily accessible with it providing the right information, proper curation from multiple sources, high modulation, assembling the data together with no redundancies, proper management, data validation, sophisticated data analysis, its integration and finally the visualization if requires etc. would all be possible with a proper background in the informatics along with the knowledge on the biological requirements. Additionally, the extreme challenges like the algorithms for the proteomic analysis demands for contemporary ways of inducing multiple sources of knowledge and high capability of analyzing the data structure. These algorithms normally would requires to define once and they should be capable on working in any kind of implementations which they are intended to work on.

4.1 PlateletWeb

The PlateletWeb knowledge base is a medium which helps investigate the signaling mechanisms and modular functions in a comprehensive manner. This functionality however is not just limited to just the hemostasis. The usability scale ranges from the pattern which can be as simple as retrieving the information about the protein of interest to the scale where the sample of proteins which are obtained in the proteomic analysis. The proteins can be visualized in the sub network format based on the interactome constructed in the PlateletWeb resource. With this the complete multi-tasking and functional modularity of the platelets can be analyzed and studied. In order to make sure the quality of the PlateletWeb resource applicable for the newest findings, I have performed the analysis on the newly identified antithrombotic targets such as LXR (*Spyridon, Moraes et al. 2011*) and on the signaling modulator PECAM1 (*Moraes, Barrett et al. 2010*). It is concluded that with the help of this resource, these targets can be easily investigated in the context of interacting partners and phosphorylation events. Additionally, a complete platelet and kinase enrichment is performed and this helped me to analyze the kinases in depth. A detailed analysis was also performed in order to see that the PlateletWeb provides a semantic, systematic and a structured overview of the platelet interactome and phosphoproteome including validated interactions and phosphorylations based on published studies and experimentally validated phosphorylation sites in platelets. Additionally, the inclusion of drugs and drug targets enabled the systemic analysis of pharmacological regulation in platelets. The complete website with the relational database as the basis enabled me to create new specific features for the PlateletWeb website which would ease the data retrieved from the PlateletWeb resource. Advanced search options in the PlateletWeb website for drugs, diseases and functional annotations effectively enable the implementation of data mining strategies for the detection of novel platelet specific targets. Additional bindings and cross links created for each and every protein helped to reduce the false positives in the resource. All proteins are also associated with their drug associations, allowing bidirectional navigation through the network of drug target relationships. This similar format was also ensured in the kinase substrate relationships. Protein kinases are key regulators of platelet signaling. The kinase activity, which is crucial in a variety of human diseases such as cancer is associated with over 400 human diseases (*Melnikova and Golden 2004*). The knowledge about human kinases and their substrates in a network context, coupled with information on associated drugs can indicate putative new pharmacological targets. My analysis provided the first insights on to the enrichment methodologies to check how and which platelet proteins play the key role. Modern phosphoproteomic studies deliver large amounts of valuable data on site-specific phosphorylation including quantitative measurements of dynamic signaling events (*Rigbolt, Prokhorova et al. 2011*). Usually, however, they offer no information about the corresponding kinase which thus remains the missing link in the reconstruction of the cellular signaling cascades from phosphoproteome data. Here, the bioinformatical predictions and literate-curated networks may provide the necessary kinase associations and deliver a cellular context for the analysis of the raw experimental data. This integrated analysis, however, is not restricted to a single protein level. The proteins of interest from a sample can be investigated at an instance by extracting a subnetwork from the human phosphoproteome, kinome and interactome and *PlateletWeb* allows the visual representation of resulting subnetworks in vector graphics or export for further analysis in Cytoscape. Various pathways considered to play a key role were extracted, visualized and analyzed for signal regulation. Alternatively, platelet proteins belonging to characteristic pathways (defined by KEGG) are highlighted and available for pathway network analysis. The enrichment analysis of drug targets and disease genes in hub platelet proteins were also performed and this resulted in the first insight which proclaimed that the highly connected proteins in the network are among the enriched drug targets and disease genes. The large amount of available platelet data gave rise to a series of analysis aimed at elucidating the systems biological background of platelet signaling. Using pathway data from KEGG the investigation is based on the enrichment for a specific pathway in platelet proteins. After testing whether the found platelet proteins are significantly higher in number than expected by chance, the test concluded that a group of pathways are found with significantly higher number of platelet proteins. Among the top-enriched pathways are key pathways in platelet signaling: cytoskeleton reorganization and focal adhesion. The *PlateletWeb* knowledge base is a comprehensive internet-based resource for platelet researchers, which presents intriguing novel options for a systems biological investigation of platelet signaling. It provides many opportunities for the investigation of dynamic network restructuring also in regard to the recently recognized condition-dependent further tasks of the platelet in infection, inflammation, cancer and sepsis (*Leslie 2010*). Future efforts in the development of *PlateletWeb* will focus on the integration of novel platelet studies (e.g. new transcriptome information from recent RNA sequencing data (*Rowley, Oler et al. 2011*)) and updated datasets from the source databases while maintaining the high standard of data curation.

4.2 Other databases (mouse and cd molecules)

4.2.1 CD Molecule database

Cluster of differentiation (CD) molecules are cell surface molecules and targets for immunophenotyping (*Fabryova and Simon 2009*). These CD molecules can act in several ways, for instace as receptors or ligands (*Gregory 2000*). Using the PlateletWeb database, a more specific database was created, termed the CD molecule database (CDDB). This is an interaction database for CD molecules, similar to the PlateletWeb database. It also provides the information about the proteins, protein information and summary, the KEGG pathways, the architeture types and tissue expression. This database gives the first insight to provide information as how the PlateletWeb database can be extended into further more specific modules. The database front end, created with PHP shows a list of total interacting proteins and a list of interacting CD molecules only. This small scale database can be used to gain information when conducting the experiments with cluster of differentiation molecules. However, similar to the PlateletWeb database, this database contains the information only on the human proteins. This might be a problem for preparing experiments as mos of them are conducted not on human beings but on the studied model organisms such as mus musculus.

Thus, this CD molecule database was expanded to the species mouse by including the orthologs and interlogs from the human proteome. The complete search functionality was also added to this database to retrieve the relavant data from the database. This database was designed by me and was helping the practical training student, Benjamin Merget for his thesis.

4.2.2 Mouse database

Mice have become a major animal model in platelet research and more and more data for this organism becomes available, the establishment of cellular signalling network of murine platelets becomes an interesting and realistic perspective. A first insight was perfomed and the data is curated manually from multiple databases (*Kanapin, Batalov et al. 2003;*

Vassilatis, Hohmann et al. 2003; Senis, Tomlinson et al. 2007; Li, Cai et al. 2010; Turner, Razick et al. 2010; Flicek, Amode et al. 2011; Rowley, Oler et al. 2011). The first database was constructed using the mouse proteome and interactome information topped up with the information on each protein and its functional information. This was then matched with the proteins from the PlateletWeb to contain the interlogs and homologs. The construction of the database, however requires lots of criteria to be considered and can be understood as a first step for the potentially high end research.

4.3 Semantic similarity

Semantic similarity measures are useful to assess the physiological relevance of protein-protein interactions. Using the annotations like Gene Ontology (GO) it is possible to quantify the similarity between the genes based on their functions. Proteins that interact in the cell are likely to be performing similar functions or involved in similar biological processes or even found on the same component. Here an algorithm was introduced to extract the functional modules from the proteins identified in the proteomic analysis using the semantic similarity and the exact approach (heinz). The GOSim package (*Frohlich, Speer et al. 2007*) was used in order to calculate the semantic similarity between the proteins. The GOSim package offers an easy way to gain insights to the functional modules of the genes.

4.3.1 Semantic Similarity of Terms

The semantic similarity between the terms has to be considered before moving onto the similarity between the genes. Various methods (*Resnik 1995; Lord, Stevens et al. 2003; Schlicker, Domingues et al. 2006; Wang, Du et al. 2007; Schlicker and Albrecht 2008; Benabderrahmane, Smail-Tabbone et al. 2010; Yu, Li et al. 2010*) are available to calculate the semantic similarity between two terms. The usage of the semantic similarity was discussed thoroughly in multiple reviews (*Guo, Liu et al. 2006; Xu, Du et al. 2008; Pesquita, Faria et al. 2009*), and Resnik was considered the best applicable measure for GO based semantic similarity on PPI.

Semantic similarity measures can be broadly classified into two groups, the edge based and the node based. The similarity based on the shared paths between two terms is given by the edge based (*Cheng, Cline et al. 2004; Wu, Su et al. 2005; Yu, Gao et al. 2005; del Pozo, Pazos et al. 2008*) methods. The node based methods (*Resnik 1995; Jiang and Conrath 1997; Lin 1998*) on the other hand rely on comparing the properties of the input terms taking their terms themselves, ancestors and the descendants into account. Information content (IC) is the concept which is commonly used in all these approaches. The IC gives a measure of how specific and informative the term is. The probability assigned to a term is defined as its relative frequency of occurrence. The root has probability *p(root) = 1* if it is unique. The IC of a term c can be quantified as the negative log likelihood,

$$-\,log\ p(c)$$

where p(c) is the probability of occurrence of c in a specific corpus being normally estimated by its frequency of annotation. This is also called as most informative common ancestor (MICA).

IC provides a measure of a terms specificity that is independent of its depth in the Ontology. This is due to the fact that the IC of a term is dependent on its children but not on its parents. Because of this main reason, the approaches based on IC are less sensitive to the issues of variable semantic distance and variable node density. However, IC is biased because terms related to areas of scientific interest are expected to be well annotated when compared to other terms. Even in such cases, the use of the IC still makes sense from a probabilistic point of view, as it is more probable and less meaningful that two gene products share a commonly used term than an uncommonly used term(*Pesquita, Faria et al. 2009*).

Resnik measure

Resnik (*Resnik 1995*) uses the concept of information content “IC” to define more consistent semantic similarity measure. The similarity between two terms is high if the information shared between them is higher. This shared information is considered by the set of common ancestors in the graph. The amount of shared information and thus the similarity between the two terms is quantified by the information content of the common ancestors. Thus, the semantic similarity between two terms can be given as

$$\mathrm{sim}_{\mathrm{Resnik}}(\mathrm{c}_1, \mathrm{c}_2) \; = \max_{c \in s(c_1, c_2)} (-\log p(c))$$

where $s(c_1, c_2)$ is the set of common ancestors of terms c_1 and c_2.

The above equation can also be written as:

$$\mathrm{sim}_{\mathrm{Resnik}}(\mathrm{c}_1, \mathrm{c}_2) \; = \max_{c \in s(c_1, c_2)} \mathrm{IC}(c_{MICA})$$

In this measure, the minimum similarity is zero and there is no maximum for this measure.

"Lin's" and "Jiang and Conrath's" Measure

Resnik measure is effective in determining the information shared by two terms; however it does not consider how distant the terms are from their common ancestors. Lin (*Lin 1998*) then defines the similarity between two terms as the ratio of the distance of terms and the information needed to fully describe the two terms. The commonality of the terms can be defined by the common ancestors of the terms. To take that distance into account, both Lin's and Jiang and Conrath's (*Jiang and Conrath 1997*) measure relate the IC of the MICA to the IC of the terms being compared. The information needed to fully describe both terms is the sum of their information, since the random selection of one term is independent of the random selection of the second term. The defining equation is given by

$$\mathrm{Sim}_{\mathrm{Lin}}(\mathrm{c}_1, \mathrm{c}_2) = \frac{2\ x\ IC(cMICA)}{IC(c1)+IC(c2)}$$

$$\mathrm{Sim}_{\mathrm{JC}}(\mathrm{c}_1, \mathrm{c}_2) = 1 - \mathrm{IC}(\mathrm{c}_1) + \mathrm{IC}(\mathrm{c}_2) - 2 \text{ x } \mathrm{IC}(\mathrm{c}_{\mathrm{MICA}})$$

In both the cases, $S(c_1, c_2)$ are the sets of common ancestors of terms c_1 and c_2. The minimum possible similarity value in these methods is 0 and the maximum is 1.

However, being relative measures, similarity measures of both Lin and JC are displaced from the graph. The measures are proportional to the IC differences between the terms and their common ancestor, independently of the absolute IC of the ancestor. The relevance is not considered in both the measures.

Schlicker Measure

To overcome this limitation, Schlicker (*Schlicker, Domingues et al. 2006*) tuned up with a measure which would combine both Lin and Resnik similarity measures. This would help to take the relevance information into account, which takes the lowest common ancestor into account. The probability of the lowest common ancestor reflects the level of detail. Generic terms do not have a high relevance for the comparison of the exact function of different gene products. The Schlicker measure is as follows:

$$\mathrm{Sim}_{\mathrm{Rel}}(\mathrm{c}_1, \mathrm{c}_2) \;=\; \max_{c \in s(c_1, c_2)} \left(\frac{2\,.\log p(c)}{\log p(c1) + \log p(c2)} \left(1 - p(c)\right) \right)$$

This equation can as well be written as:

$$\mathrm{Sim}_{\mathrm{Rel}}(\mathrm{c}_1,\mathrm{c}_2) = \mathrm{Sim}_{\mathrm{Lin}}(\mathrm{c}_1,\mathrm{c}_2) \text{ x } (1\text{-p}(\mathrm{c}_{\mathrm{MICA}}))$$

Similar to Lin, Rel measure also has the minimum similarity value of 0 and the maximum similarity value of 1. The relevance of a term decreases with the increasing probability in this method.

4.3.2 Semantic Similarity between Genes

Genes are usually annotated with more than one gene ontology terms in the GO database. In this scenario, in order to calculate the semantic similarity between the two genes, the semantic similarity between all the GO terms of each gene should be compared. Computation of maximum and average similarity between any pair of GO terms is defined in the existing tools like FuSSiMeg (*Couto, Silva et al. 2003*), eGOn (*GÜnther, Langaas et al. 2006*) and DAVID (*Huang da, Sherman et al. 2009; Huang da, Sherman et al. 2009*). However, eGOn and DAVID approaches measure gene functional similarities based on the probability of the appearance of GO terms or the kappa statistics of similar annotation terms correlated with different genes and ignore the semantic relations (is-a and part-of) among these terms in the GO graph.

Functional Semantic Similarity Measures between Gene-Products (FuSSiMeG) is a functional similarity measure between gene-products that compares the semantic similarity between the terms in their annotations. This is a hybrid semantic similarity measure, which integrates the information content with conceptual distance (edge based) factors. In order to compute the similarity between functional properties, FuSSiMeG assumes that two gene-products have a functional similarity when they are annotated with similar functional terms. FuSSiMeG then measures the similarity between gene products by the maximum similarity between their assigned terms.

Given two terms *g* and *g´* annotated with GO terms t_1, t_2 … t_n and $t_1´$, $t_2´$… $t_m´$ the functional similarity between two genes *g* and *g´* is defined as

$$\text{Sim gene}(g,g´) = \max_{\substack{i=1,\dots,n \\ j=1,\dots,m}} sim(t_i,t_j´)$$

Here the semantic similarity measure ranges between 0 and 1.

Another semantic similarity measure, G-SESAME (*Wang, Du et al. 2007*) addresses the critical need of determining the functional similarities based on gene annotation information from heterogeneous data sources. This algorithm focuses on the relationships of the GO terms with in a specific ontology to determine the semantic similarity which helps in gaining the consistent measurement between two GO terms.

Another graph based method is GOSemSim (*Yu, Li et al. 2010*), which is an R package for measuring the semantic similarity among GO terms and gene products. Many functions are provided by the GOSemSim package, *geneSim* is specifically used to compute the semantic similarity between GO descriptions of gene products. Different measures can be used along with this function, some of them being “Wang”, "Resnik", "Lin", "Rel", and "Jiang". However when this package is tested on the human species for the complete interactome with multiple measures, the output has extremely biased results near the minimum and the maximum. When two similar genes are compared, it would give a 1 as the output value. Many gene products are forming complexes which perform the same biological function and therefore have almost similar functionality. However, in order to identify the differences in similar genes (not the same gene); the focus has to be shifted to similar method providing higher resolution.

In this regard, a complementary tool which beholds similar functionality is GOSim (*Frohlich, Speer et al. 2007*). Taken FuSSiMeg as basic idea, GOSim extends the resulting value by further normalizing to account for an unequal number of GO terms for both genes.

$$\text{Sim gene}(g,g') = \frac{sim_{gene}(g,g')}{\sqrt{sim_{gene}(g,g)sim_{gene}(g'g')}}$$

The GOSim package systematically integrates existing tools like FuSSiMeg by integrating its functionality and providing additional similarity concepts for gene products. This is implemented as a package for statistical computing environment R and has been integrated into the CRAN project.

4.3.3 Unique score between two genes

The usage of GOSim package helped to gain the semantic similarity between two genes using the function getGeneSim. However, the Gene Ontology is defined in three broad categories – Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). The Semantic similarity has to be calculated for given two genes on all the three ontologies considering one at a time. This would result in three scores for the given two genes. For instance, when quantifying the semantic similarity between the genes “VASP” (Entrez gene identifier – 7408) and “SRC” (Entrez gene identifier – 6714) the GOSim package returns the semantic similarity for BP as 0.4608678, CC as 0.3648075 and MF as 0.2174619. Schlicker et al (*Schlicker, Domingues et al. 2006*) introduced *funsim* score from the scores of BP and

MF of a pair of gene products. Two gene products with a high score in one ontology but an average score in the other can be considered average matches. However, in such case, their score should be higher than the score of two gene products that are average matches in both categories. Here, simply adding the two scores or taking the average scores would not distinguish between these two cases. In order to gain the distinction, the squaring of the scores favors high similarity in one and a low score in the other one over average scores in both ontologies. The *funSim* score ranges between 0 for completely unrelated gene products and 1 for gene products with identical functionality. Due to its definition, the *funSim* score is lower than the average two scores in most cases. In order to obtain a more intuitive score, the *rfunSim* score (*Schlicker, Rahnenfuhrer et al. 2007*) was defined for two gene products were defined. However, although the three ontologies scores were considered, I stick to the *funSimAll*, with all the three scores considered together, once again defined by Schlicker et al [http://funsimmat.bioinf.mpi-inf.mpg.de/help3.php].

The funSimAll score is calculated from the Scores of BP, MF and CC for a pair of genes. It is defined as:

$$funSimAll(p,q) = \frac{1}{3}\left[\left(\frac{BPScore(p,q)}{max_{BPscore}}\right)^2 + \left(\frac{MFScore(p,q)}{max_{MFscore}}\right)^2 + \left(\frac{CCScore(p,q)}{max_{CCscore}}\right)^2\right]$$

In this equation, max(BPscore), max(MFscore) and max(CCscore) denotes the maximum possible score for BP, MF and CC respectively. Once again, the *funSimAll* score ranges between 0 for completely unrelated gene products and 1 for gene products with identical functionality. Introducing the semantic similarity calculated between two genes VASP and SRC, the result would be

$$funSimAll(p,q) = \frac{1}{3}\left[\left(\frac{0.4608678}{1}\right)^2 + \left(\frac{0.2174619}{1}\right)^2 + \left(\frac{0.3648075}{1}\right)^2\right]$$

The funSimAll result of this equation is 0.130924. However the proteins are highly annotated in the Biological Process as the interacting proteins share the common biological goal and they are comparatively less in both Molecular Function and Cellular Component. This would directly affect the *funSimAll* calculation, if the annotations are missing for one or more of the ontologies for the set of proteins. Prominent genes might lose functional information if they are not annotated in all 3 ontologies. In order to overcome this issue, the GOScores are converted into the p-values and these are then aggregated using the aggregate statistic based on the distribution of the order statistics (*Dittrich, Klau et al. 2008; Beisser, Klau et al. 2010*). Thus a single functional edge score is achieved for each of the gene pairs.

5 Conclusion and Outlook

The systems biology insights are possible due to the cleaver construction of the database. The PlateletWeb is a systems biological workbench, is a valuable resource which provides a much needed complete information for the analysis of platelet signaling in the functional context of integrated networks. The establishment of this database provides trivial and vital information about the platelet proteome and the interactome when analyzing the protein alone or when considering its role in the network context. The information gained from multiple resources and databases, improves the accuracy and the consistency in the data and decreases the false positives if any. In general, meeting the highly relevant and computational challenges of systems biology regards for powerful biological database which was then developed to knowledgebase. The functionality of the website with its advanced features like the "advanced search" provides the information only specific to the context. Additional information about the protein physical attributes, the Gene ontology information about each of the protein, the drug target information and the disease associations provide an ample function tools for platelet signaling analysis. The graphical visualizations available in the PlateletWeb knowledgebase helps understand the protein from the context of the network modules. The kinase, phosphatase information provides the signaling flow in the network context and provides ample of information for complete analysis of the modules. PlateletWeb allows detailed topological, interactome and phosphoproteome analysis and thus serves as a valuable platform for the comprehensive cellular network analysis, for instance regarding PGI2 and ADP P2Y12 receptor signaling pathways.

The analysis on the interactome network provides new insights on how the data is distributed across the network. Also, the platelet specific kinase tree revealed the distribution and enrichment of Tyrosine kinase substrates in the human platelets. The motif analysis gives the insights on how the information is processed in the anucleate cell types. The unique patterns of interactions between proteins that appear significantly more often in the real network compared with randomized networks were identified, thus gaining the new insights on the types of network regulation. The 8 key motifs with high biological importance and highest z-values are identified. The complete detailed analysis was performed to understand these key motifs and the way the motif 4574 play a key role in the network regulation. Motif analysis has further increased the knowledge of what possible mechanisms might be involved in the platelet phosphoproteome and interactome network and identifying the most important motifs in the platelet interactome. The availability of Gene Ontology information for most of the proteins helped in understanding the proteins in terms of their biological process, molecular functionality and the cellular component. This information is used for the complete gene enrichment analysis which provides the insights for how the proteins are scattered and grouped together. Taking this Gene ontology information about each of the proteins, the quantification of the relation between two proteins was achieved. This was then introduced to the heinz algorithm resulting in the biologically significant modules for the first time. This quantification, performed by taking the gene ontology information between two proteins which are interacting between each other and then checking for the semantic similarity between them. This helped to extract and understand the modules from the complete network of proteins which are specific to a certain criteria. A complete testing was performed by taking already existing modules and also by investigating new modules in different proteins identified in multiple samples.

The database can be further extended and this was also performed by extracting only the CD molecules and identifying the components. Using the PlateletWeb database, a more specific database was created, termed the CD molecule database (CDDB). This database has the capability to expand to multiple species and this was clearly concluded on a high note with the introduction of mouse interactome data. The orthologs and the interlogs of the mouse and human data can bring the complete analysis of the protein with focusing on multiple aspects. The resource can be paired up with multiple species, which for obvious reasons can be a valuable asset to the field of science and medicine.

6 Supplementary material

Supplementary Table 1 - Information on Gene Ontology enrichment analysis for Motifs 110 (5a), 4574 (5b) and 6558 (5c). Biological process (BP), molecular function (MF) and cellular component (CC) analysis are presented with the corrected p-value and the number of proteins from the real network (=biological network), which are associated with this GO-term. The p-values are adjusted using the Benjamini and Hochberg correction(*Benjamini and Yekutieli 2001*).

Abbreviations: RN - Number of proteins from real network; NPGT - Number of proteins from GO-term;

Supplementary Table 1a – Biological Process; Motif 110:

Total number of proteins in the network: 300; Number of all GO annotated Proteins: 8652

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
8286	4.53E-12	3.49E-10	12	24	Insulin receptor signaling pathway
7169	3.16E-11	2.24E-09	27	172	transmembrane receptor protein tyrosine kinase signaling pathway
43066	5.94E-10	3.64E-08	27	195	negative regulation of apoptosis
43069	7.50E-10	4.44E-08	27	197	negative regulation of programmed cell death
1775	2.33E-08	1.12E-06	21	145	cell activation
30036	6.08E-08	2.77E-06	21	153	actin cytoskeleton organization and biogenesis
30029	6.92E-08	3.07E-06	22	168	actin filament-based process

GO ID	***p-value***	***Corrected p-value***	***RN***	***NPGT***	***Description of GO Term***
45321	1.07E-07	4.64E-06	18	118	leukocyte activation
7265	2.86E-07	1.16E-05	15	88	Ras protein signal transduction
51049	8.48E-07	3.14E-05	18	135	regulation of transport
32879	9.48E-07	3.43E-05	18	136	regulation of localization

Supplementary Table 1b – Molecular Function; Motif 110:

Total number of proteins in the network: 308; Number of all GO annotated Proteins: 9108

GO ID	***p-value***	***Corrected p-value***	***RN***	***NPGT***	***Description of GO Term***
4715	3.07E-16	2.65E-14	17	36	non-membrane spanning protein tyrosine kinase activity
4713	4.80E-15	2.48E-13	30	156	protein tyrosine kinase activity
19904	3.08E-11	9.35E-10	23	127	protein domain specific binding
4697	8.18E-09	2.11E-07	8	15	protein kinase C activity
5057	9.04E-09	2.23E-07	20	128	receptor signaling protein activity
4716	3.24E-08	7.61E-07	7	12	receptor signaling protein tyrosine kinase activity
32403	2.22E-06	4.25E-05	14	92	protein complex binding

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
5158	5.20E-06	9.60E-05	7	22	insulin receptor binding
43560	6.24E-06	1.11E-04	4	5	insulin receptor substrate binding
19992	1.08E-05	1.74E-04	10	54	diacylglycerol binding
43548	1.82E-05	2.86E-04	4	6	phosphoinositide 3-kinase binding

Supplementary Table 1c – Cellular Component; Motif 110:

Total number of proteins in the network: 298; Number of all GO annotated Proteins: 9082

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
5912	1.36E-08	6.96E-07	12	45	adherens junction
5925	1.56E-07	6.32E-06	9	28	focal adhesion
31252	6.20E-07	1.96E-05	12	62	leading edge
5924	5.64E-07	1.96E-05	9	32	cell-substrate adherens junction
45121	1.07E-06	2.77E-05	11	54	membrane raft
30055	9.95E-07	2.77E-05	9	34	cell-substrate junction
48770	5.90E-06	1.29E-04	12	76	pigment granule
42470	5.90E-06	1.29E-04	12	76	melanosome
16323	1.16E-05	1.94E-04	12	81	basolateral plasma membrane
5942	3.81E-05	5.15E-04	5	13	phosphoinositide 3-kinase complex

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
1726	7.82E-05	1.01E-03	7	33	ruffle

Supplementary Table 1d – Biological Process; Motif 4574:

Total number of proteins in the network: 158; Number of all GO annotated Proteins: 8652

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
8286	1.08E-13	1.06E-11	11	24	insulin receptor signaling pathway
7169	4.23E-13	3.86E-11	22	172	transmembrane receptor protein tyrosine kinase signaling pathway
7265	5.64E-09	2.67E-07	13	88	Ras protein signal transduction
43066	8.43E-08	3.59E-06	17	195	negative regulation of apoptosis
43069	9.79E-08	3.79E-06	17	197	negative regulation of programmed cell death
7159	2.32E-07	8.49E-06	6	16	leukocyte adhesion
46777	5.93E-07	2.05E-05	8	41	protein amino acid autophosphorylation
16540	8.73E-07	2.74E-05	8	43	protein autoprocessing
48009	3.59E-06	1.07E-04	4	7	insulin-like growth factor receptor signaling pathway
30856	3.59E-06	1.07E-04	4	7	regulation of epithelial cell differentiation

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
51049	6.04E-06	1.64E-04	12	135	regulation of transport

Supplementary Table 1e – Molecular Function; Motif 4574:

Total number of proteins in the network: 162; Number of all GO annotated Proteins: 9108

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
4715	1.18E-17	7.23E-16	15	36	non-membrane spanning protein tyrosine kinase activity
4713	3.30E-14	1.73E-12	22	156	protein tyrosine kinase activity
4716	2.44E-08	5.27E-07	6	12	receptor signaling protein tyrosine kinase activity
32403	7.07E-08	1.37E-06	12	92	protein complex binding
5158	6.74E-08	1.37E-06	7	22	insulin receptor binding
4697	1.27E-07	2.32E-06	6	15	protein kinase C activity
5057	3.98E-07	6.35E-06	13	128	receptor signaling protein activity
43560	4.76E-07	7.06E-06	4	5	insulin receptor substrate binding
43548	1.41E-06	1.91E-05	4	6	phosphoinositide 3-kinase binding
19904	2.45E-06	3.21E-05	12	127	protein domain specific binding
42169	8.71E-06	1.07E-04	5	17	SH2 domain binding

Supplementary Table 1f – Cellular Component; Motif 4574:

Total number of proteins in the network: 159; Number of all GO annotated Proteins: 9082

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
45121	2.52E-08	1.17E-06	10	54	membrane raft
5912	9.20E-07	3.06E-05	8	45	adherens junction
16599	4.03E-06	1.04E-04	5	15	caveola
5916	5.13E-05	7.97E-04	3	5	fascia adherens
30018	7.90E-05	1.15E-03	4	14	Z disc
31252	9.68E-05	1.33E-03	7	62	leading edge
5925	1.10E-04	1.42E-03	5	28	focal adhesion
5924	2.12E-04	2.47E-03	5	32	cell-substrate adherens junction
30055	2.85E-04	3.02E-03	5	34	cell-substrate junction
48770	3.51E-04	3.28E-03	7	76	pigment granule
42470	3.51E-04	3.28E-03	7	76	melanosome

Supplementary Table 1g – Biological Process; Motif 6558:

Total number of proteins in the network: 90; Number of all GO annotated Proteins: 8652

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
8286	1.09E-12	5.72E-11	9	24	insulin receptor signaling pathway
7169	1.42E-12	7.11E-11	17	172	transmembrane receptor protein tyrosine kinase signaling pathway
46777	7.22E-09	2.65E-07	8	41	protein amino acid autophosphorylation
43066	1.10E-08	3.79E-07	14	195	negative regulation of apoptosis
16540	1.08E-08	3.79E-07	8	43	protein autoprocessing
43069	1.26E-08	4.19E-07	14	197	negative regulation of programmed cell death
165	3.75E-07	9.85E-06	10	119	MAPKKK cascade
48009	3.74E-07	9.85E-06	4	7	insulin-like growth factor receptor signaling pathway
30856	3.74E-07	9.85E-06	4	7	regulation of epithelial cell differentiation
43405	5.37E-07	1.36E-05	9	95	regulation of MAP kinase activity
16485	7.74E-07	1.85E-05	8	73	protein processing

Supplementary Table 1h – Molecular Function; Motif 6558:

Total number of proteins in the network: 95; Number of all GO annotated Proteins: 9108

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
4715	3.23E-23	2.09E-21	16	36	non-membrane spanning protein tyrosine kinase activity
4713	7.86E-17	3.39E-15	20	156	protein tyrosine kinase activity
5057	5.93E-10	9.03E-09	13	128	receptor signaling protein activity
4716	9.65E-10	1.39E-08	6	12	receptor signaling protein tyrosine kinase activity
4697	5.09E-09	6.28E-08	6	15	protein kinase C activity
19904	6.56E-09	7.72E-08	12	127	protein domain specific binding
43560	5.51E-08	5.95E-07	4	5	insulin receptor substrate binding
19992	7.29E-08	7.56E-07	8	54	diacylglycerol binding
42169	6.22E-07	6.20E-06	5	17	SH2 domain binding
5158	2.54E-06	2.35E-05	5	22	insulin receptor binding
4715	3.23E-23	2.09E-21	16	36	insulin-like growth factor receptor binding

Supplementary Table 1i – Cellular Component; Motif 6558:

Total number of proteins in the network: 92; Number of all GO annotated Proteins: 9082

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
45121	1.04E-06	5.78E-05	7	54	membrane raft

GO ID	*p-value*	*Corrected p-value*	*RN*	*NPGT*	*Description of GO Term*
16599	1.24E-05	4.10E-04	4	15	caveola
5912	8.48E-05	2.17E-03	5	45	adherens junction
31252	3.92E-04	8.14E-03	5	62	leading edge
30877	6.01E-04	1.11E-02	2	4	beta-catenin destruction complex
5916	9.95E-04	1.27E-02	2	5	fascia adherens
16323	1.34E-03	1.58E-02	5	81	basolateral plasma membrane
5925	2.74E-03	2.53E-02	3	28	focal adhesion
5924	4.03E-03	3.52E-02	3	32	cell-substrate adherens junction
30027	5.21E-03	3.97E-02	3	35	lamellipodium
30055	4.80E-03	3.97E-02	3	34	cell-substrate junction

7 References

(2008). "The Gene Ontology project in 2008." Nucleic Acids Res **36**(Database issue): D440-444.

(2009). "The Universal Protein Resource (UniProt) 2009." Nucleic Acids Res **37**(Database issue): D169-174.

Alfarano, C., C. E. Andrade, et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update.

Alon, U. (2003). "Biological networks: the tinkerer as an engineer." Science **301**(5641): 1866-1867.

Alonso, A., J. Sasin, et al. (2004). "Protein tyrosine phosphatases in the human genome." Cell **117**(6): 699-711.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Audenaert, P., T. Van Parys, et al. (2011). "CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks." Bioinformatics **27**(11): 1587-1588.

Batagelj, V. and A. Mrvar (1998). "Pajek – Program for Large Network Analysis." Connections **21**: 1-11.

Batagelj, V. and A. Mrvar (2003). "PAJEK ANALYSIS AND VISUALIZATION OF LARGE NETWORKS." Science **2265**: 8-11.

Beisser, D., G. W. Klau, et al. (2010). "BioNet: an R-Package for the functional analysis of biological networks." Bioinformatics **26**(8): 1129-1130.

Benabderrahmane, S., M. Smail-Tabbone, et al. (2010). "IntelliGO: a new vector-based semantic similarity measure including annotation origin." BMC Bioinformatics **11**: 588.

Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." Annals of Statistics **29**: 1165--1188.

Boyanova, D., S. Nilla, et al. (2012). "PlateletWeb: a systems biologic analysis of signaling networks in human platelets." Blood **119**(3): e22-34.

Brill, L. M., W. Xiong, et al. (2009). "Phosphoproteomic analysis of human embryonic stem cells." Cell Stem Cell **5**(2): 204-213.

Cheng, J., M. Cline, et al. (2004). "A knowledge-based clustering algorithm driven by Gene Ontology." J Biopharm Stat **14**(3): 687-700.

Coppinger, J., D. J. Fitzgerald, et al. (2007). "Isolation of the platelet releasate." Methods Mol Biol **357**: 307--311.

Coppinger, J. A., G. Cagney, et al. (2004). "Characterization of the proteins released from activated platelets leads to localization of novel platelet proteins in human atherosclerotic lesions." Blood **103**(6): 2096--2104.

Couto, F. M., M. J. Silva, et al. (2003). "Implementation of a functional semantic similarity measure between gene-products."

De Nooy, W., A. Mrvar, et al. (2005). Exploratory social network analysis with Pajek, Cambridge University Press.

del Pozo, A., F. Pazos, et al. (2008). "Defining functional distances over gene ontology." BMC Bioinformatics **9**: 50.

Denis, M. M., N. D. Tolley, et al. (2005). "Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets." Cell **122**(3): 379-391.

Dittrich, M., I. Birschmann, et al. (2011). "Integrated platelet networks for the analysis of different system states." Current Proteomics **8**(3): 229-236.

Dittrich, M., I. Birschmann, et al. (2008). "Platelet protein interactions: map, signaling components, and phosphorylation groundstate." Arterioscler Thromb Vasc Biol **28**(7): 1326-1331.

Dittrich, M., I. Birschmann, et al. (2006). "Analysis of SAGE data in human platelets: features of the transcriptome in an anucleate cell." Thromb Haemost **95**(4): 643-651.

Dittrich, M., I. Birschmann, et al. (2005). "Understanding platelets. Lessons from proteomics, genomics and promises from network analysis." Thromb Haemost **94**(5): 916--925.

Dittrich, M., V. Strassberger, et al. (2010). "Characterization of a novel interaction between vasodilator-stimulated phosphoprotein and Abelson interactor 1 in human platelets: a concerted computational and experimental approach." Arterioscler Thromb Vasc Biol **30**(4): 843--850.

Dittrich, M. T., G. W. Klau, et al. (2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach." Bioinformatics **24**(13): i223-231.

Dittrich, M. T., G. W. Klau, et al. (2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach." Bioinformatics **24**(13): i223--i231.

Ebstein, F., P. M. Kloetzel, et al. (2012). "Emerging roles of immunoproteasomes beyond MHC class I antigen processing." Cell Mol Life Sci.

Fabryova, K. and M. Simon (2009). "Function of the cell surface molecules (CD molecules) in the reproduction processes." Gen Physiol Biophys **28**(1): 1-7.

Ferro, A., R. Giugno, et al. (2007). "NetMatch: a Cytoscape plugin for searching biological networks." Bioinformatics **23**(7): 910-912.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic Acids Res **39**(Database issue): D800-806.

Frohlich, H., N. Speer, et al. (2007). "GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products." BMC Bioinformatics **8**: 166.

García, A., S. Prabhakar, et al. (2004). "Extensive analysis of the human platelet proteome by two-dimensional gel electrophoresis and mass spectrometry." Proteomics **4**(3): 656--668.

Garcia, A., Y. A. Senis, et al. (2006). "A global proteomics approach identifies novel phosphorylated signaling proteins in GPVI-activated platelets: involvement of G6f, a novel platelet Grb2-binding membrane adapter." Proteomics **6**(19): 5332-5343.

Garcia, B. A., D. M. Smalley, et al. (2005). "The platelet microparticle proteome." J Proteome Res **4**(5): 1516-1521.

Gardiner, E. E., J. F. Arthur, et al. (2010). "GPIbalpha-selective activation of platelets induces platelet signaling events comparable to GPVI activation events." Platelets **21**(4): 244-252.

Glenister, K. M., K. A. Payne, et al. (2008). "Proteomic analysis of supernatant from pooled buffy-coat platelet concentrates throughout 7-day storage." Transfusion **48**(1): 99-107.

Goddard, A., G. Ladds, et al. (2006). "Identification of Gnr1p, a negative regulator of G alpha signalling in Schizosaccharomyces pombe, and its complementation by human G beta subunits." Fungal Genet Biol **43**(12): 840-851.

Gregory, C. D. (2000). "CD14-dependent clearance of apoptotic cells: relevance to the immune system." Curr Opin Immunol **12**(1): 27-34.

Guerrier, L., S. Claverol, et al. (2007). "Exploring the platelet proteome via combinatorial, hexapeptide ligand libraries." J Proteome Res **6**(11): 4290--4303.

GÜnther, C.-C., M. Langaas, et al. (2006). "Statistical Hypothesis Testing of Association Between Two Lists of Genes for a Given Gene Class." Cancer Research.

Guo, X., R. Liu, et al. (2006). "Assessing semantic similarity measures for the characterization of human regulatory pathways." Bioinformatics **22**(8): 967-973.

Guzzi, P. H., M. Mina, et al. (2011). "Semantic similarity analysis of protein data: assessment with biological features and issues." Brief Bioinform.

Haudek, V. J., A. Slany, et al. (2009). "Proteome maps of the main human peripheral blood constituents." J Proteome Res **8**(8): 3834-3843.

Hemerly, A. S., S. G. Prasanth, et al. (2009). "Orc1 controls centriole and centrosome copy number in human cells." Science **323**(5915): 789-793.

Hornbeck, P. V., I. Chabra, et al. (2004). "PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation." Proteomics **4**(6): 1551-1561.

Huang da, W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res **37**(1): 1-13.

Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**(5518): 929-934.

Jackson, D. E., K. R. Kupcho, et al. (1997). "Characterization of phosphotyrosine binding motifs in the cytoplasmic domain of platelet/endothelial cell adhesion molecule-1 (PECAM-1) that are required for the cellular association and activation of the protein-tyrosine phosphatase, SHP-2." J Biol Chem **272**(40): 24868--24875.

Jackson, D. E., C. M. Ward, et al. (1997). "The protein-tyrosine phosphatase SHP-2 binds platelet/endothelial cell adhesion molecule-1 (PECAM-1) and forms a distinct signaling complex during platelet aggregation. Evidence for a mechanistic link between PECAM-1- and integrin-mediated cellular signaling." J Biol Chem **272**(11): 6986-6993.

Jiang, J. J. and D. W. Conrath (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.

Kanapin, A., S. Batalov, et al. (2003). "Mouse proteome analysis." Genome Res **13**(6B): 1335-1344.

Kashtan, N., S. Itzkovitz, et al. (2004). "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs." Bioinformatics **20**(11): 1746-1758.

Keshava Prasad, T. S., R. Goel, et al. (2009). "Human Protein Reference Database--2009 update." Nucleic Acids Res **37**(Database issue): D767-772.

Knox, C., V. Law, et al. (2011). "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic Acids Res **39**(Database issue): D1035-1041.

Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.

Lange, A., S. A. Wickström, et al. (2009). "Integrin-linked kinase is an adaptor with essential functions during mouse development." Nature **461**(7266): 1002--1006.

Lee, K. Y., S. W. Bang, et al. (2012). "Phosphorylation of ORC2 protein dissociates origin recognition complex from chromatin and replication origins." J Biol Chem **287**(15): 11891-11898.

Leslie, M. (2010). "Cell biology. Beyond clotting: the powers of platelets." Science **328**(5978): 562-564.

Letunic, I. and P. Bork (2007). "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation." Bioinformatics **23**(1): 127-128.

Lewandrowski, U., S. Wortelkamp, et al. (2009). "Platelet membrane proteomics: a novel repository for functional research." Blood **114**(1): e10-19.

Li, X., H. Cai, et al. (2010). "A mouse protein interactome through combined literature mining with multiple sources of interaction evidence." Amino Acids **38**(4): 1237-1252.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann.

Linding, R., L. J. Jensen, et al. (2007). "Systematic discovery of in vivo phosphorylation networks." Cell **129**(7): 1415-1426.

Liu, B. A., K. Jablonowski, et al. (2006). "The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling." Mol Cell **22**(6): 851-868.

Liu, N., W. Song, et al. (2008). "Proteomics analysis of differential expression of cellular proteins in response to avian H9N2 virus infection in human cells." Proteomics **8**(9): 1851-1858.

Ljubic, I., R. Weiskircher, et al. (2006). "An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem." Mathematical Programming **105**(2-3): 427-449.

Lord, P. W., R. D. Stevens, et al. (2003). "Semantic similarity measures as tools for exploring the gene ontology." Pac Symp Biocomput: 601-612.

Maglott, D., J. Ostell, et al. (2007). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **35**(Database issue): D26-31.

Mangan, S. and U. Alon (2003). "Structure and function of the feed-forward loop network motif." Proc Natl Acad Sci U S A **100**(21): 11980-11985.

Manning, G., D. B. Whyte, et al. (2002). "The protein kinase complement of the human genome." Science **298**(5600): 1912-1934.

Marcus, K., D. Immler, et al. (2000). "Identification of platelet proteins separated by two-dimensional gel electrophoresis and analyzed by matrix assisted laser desorption/ionization-time of flight-mass spectrometry and detection of tyrosine-phosphorylated proteins." Electrophoresis **21**(13): 2622--2636.

Martens, L., P. Van Damme, et al. (2005). "The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile." Proteomics **5**(12): 3193-3204.

Maynard, D. M., H. F. G. Heijnen, et al. (2007). "Proteomic analysis of platelet alpha-granules using mass spectrometry." J Thromb Haemost **5**(9): 1945--1955.

Melnikova, I. and J. Golden (2004). "Targeting protein kinases." Nat Rev Drug Discov **3**(12): 993-994.

Mendez, J., X. H. Zou-Yang, et al. (2002). "Human origin recognition complex large subunit is degraded by ubiquitin-mediated proteolysis after initiation of DNA replication." Mol Cell **9**(3): 481-491.

Miller, M. L., L. J. Jensen, et al. (2008). "Linear motif atlas for phosphorylation-dependent signaling." Sci Signal **1**(35): ra2.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-827.

Moebius, J., R. P. Zahedi, et al. (2005). "The human platelet membrane proteome reveals several new potential membrane proteins." Mol Cell Proteomics **4**(11): 1754--1761.

Moraes, L. A., N. E. Barrett, et al. (2010). "Platelet endothelial cell adhesion molecule-1 regulates collagen-stimulated platelet function by modulating the association of phosphatidylinositol 3-kinase with Grb-2-associated binding protein-1 and linker for activation of T cells." J Thromb Haemost **8**(11): 2530-2541.

Newman, P. J. and D. K. Newman (2003). "Signal transduction pathways mediated by PECAM-1: new roles for an old molecule in platelet and vascular cell biology." Arterioscler Thromb Vasc Biol **23**(6): 953-964.

Nikolopoulos, S. N. and C. E. Turner (2001). "Integrin-linked kinase (ILK) binding to paxillin LD1 motif regulates ILK localization to focal adhesions." J Biol Chem **276**(26): 23499-23505.

O'Neill, E. E., C. J. Brock, et al. (2002). "Towards complete analysis of the platelet proteome." Proteomics **2**(3): 288--305.

Olsen, J. V., B. Blagoev, et al. (2006). "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks." Cell **127**(3): 635--648.

Peri, S., J. D. Navarro, et al. (2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans." Genome Res **13**(10): 2363-2371.

Pesquita, C., D. Faria, et al. (2009). "Semantic similarity in biomedical ontologies." PLoS Comput Biol **5**(7): e1000443.

Piersma, S. R., H. J. Broxterman, et al. (2009). "Proteomics of the TRAP-induced platelet releasate." J Proteomics **72**(1): 91-109.

Pounds, S. and S. W. Morris (2003). "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values." Bioinformatics **19**(10): 1236-1242.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.

Rigbolt, K. T., T. A. Prokhorova, et al. (2011). "System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation." Sci Signal **4**(164): rs3.

Rowley, J. W., A. Oler, et al. (2011). "Genome wide RNA-seq analysis of human and mouse platelet transcriptomes." Blood.

Schlicker, A. and M. Albrecht (2008). "FunSimMat: a comprehensive functional similarity database." Nucleic Acids Res **36**(Database issue): D434-439.

Schlicker, A., F. S. Domingues, et al. (2006). "A new measure for functional similarity of gene products based on Gene Ontology." BMC Bioinformatics **7**: 302.

Schlicker, A., J. Rahnenfuhrer, et al. (2007). "GOTax: investigating biological processes and biochemical activities along the taxonomic tree." Genome Biol **8**(3): R33.

Schreiber, F. and H. Schwobbermeyer (2005). "MAVisto: a tool for the exploration of network motifs." Bioinformatics **21**(17): 3572-3574.

Schwertz, H., N. D. Tolley, et al. (2006). "Signal-dependent splicing of tissue factor pre-mRNA modulates the thrombogenicity of human platelets." J Exp Med **203**(11): 2433-2440.

Schwobbermeyer, H. and R. Wunschiers (2012). "MAVisto: a tool for biological network motif analysis." Methods Mol Biol **804**: 263-280.

Senis, Y. A., M. G. Tomlinson, et al. (2007). "A comprehensive proteomics and genomics analysis reveals novel transmembrane proteins in human platelets and mouse megakaryocytes including G6b-B, a novel immunoreceptor tyrosine-based inhibitory motif protein." Mol Cell Proteomics **6**(3): 548--564.

Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.

Springer, D. L., J. H. Miller, et al. (2009). "Platelet proteome changes associated with diabetes and during platelet storage for transfusion." J Proteome Res **8**(5): 2261--2272.

Spyridon, M., L. A. Moraes, et al. (2011). "LXR as a novel antithrombotic target." Blood **117**(21): 5751-5761.

Stark, C., B. J. Breitkreutz, et al. (2011). "The BioGRID Interaction Database: 2011 update." Nucleic Acids Res **39**(Database issue): D698-704.

Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-539.

Thiele, T., L. Steil, et al. (2007). "Profiling of alterations in platelet proteins during storage of platelet concentrates." Transfusion **47**(7): 1221-1233.

Thon, J. N., P. Schubert, et al. (2008). "Comprehensive proteomic analysis of protein changes during platelet storage requires complementary proteomic approaches." Transfusion **48**(3): 425-435.

Tunon, J., J. L. Martin-Ventura, et al. (2010). "Proteomic strategies in the search of new biomarkers in atherothrombosis." J Am Coll Cardiol **55**(19): 2009-2016.

Turner, B., S. Razick, et al. (2010). "iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence." Database (Oxford) **2010**: baq023.

Varga-Szabo, D., I. Pleines, et al. (2008). "Cell adhesion mechanisms in platelets." Arterioscler Thromb Vasc Biol **28**(3): 403-412.

Vassilatis, D. K., J. G. Hohmann, et al. (2003). "The G protein-coupled receptor repertoires of human and mouse." Proc Natl Acad Sci U S A **100**(8): 4903-4908.

Wang, J. Z., Z. Du, et al. (2007). "A new method to measure the semantic similarity of GO terms." Bioinformatics **23**(10): 1274-1281.

Wang, S., N. Nath, et al. (1999). "Rb and prohibitin target distinct regions of E2F1 for repression and respond to different upstream signals." Mol Cell Biol **19**(11): 7447-7460.

Wernicke, S. and F. Rasche (2006). "FANMOD: a tool for fast network motif detection." Bioinformatics **22**(9): 1152-1153.

Weyrich, A. S., H. Schwertz, et al. (2009). "Protein synthesis by platelets: historical and new perspectives." J Thromb Haemost **7**(2): 241-246.

Wong, J. W. H., J. P. McRedmond, et al. (2009). "Activity profiling of platelets by chemical proteomics." Proteomics **9**(1): 40--50.

Wu, H., Z. Su, et al. (2005). "Prediction of functional modules based on comparative genome analysis and Gene Ontology application." Nucleic Acids Res **33**(9): 2822-2837.

Wu, X., S. J. Noh, et al. (1996). "Selective activation of MEK1 but not MEK2 by A-Raf from epidermal growth factor-stimulated Hela cells." J Biol Chem **271**(6): 3265-3271.

Xu, T., L. Du, et al. (2008). "Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data." BMC Bioinformatics **9**: 472.

Yang, Y. X., Z. Q. Xiao, et al. (2006). "Proteome analysis of multidrug resistance in vincristine-resistant human gastric cancer cell line SGC7901/VCR." Proteomics **6**(6): 2009-2021.

Yu, G., F. Li, et al. (2010). "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products." Bioinformatics **26**(7): 976-978.

Yu, H., L. Gao, et al. (2005). "Broadly predicting specific gene functions with expression similarity and taxonomy similarity." Gene **352**: 75-81.

Yu, Y., T. Leng, et al. (2010). "Global analysis of the rat and human platelet proteome - the molecular blueprint for illustrating multi-functional platelets and cross-species function evolution." Proteomics **10**(13): 2444-2457.

Zahedi, R. P., U. Lewandrowski, et al. (2008). "Phosphoproteome of resting human platelets." J Proteome Res **7**(2): 526--534.

Zaidel-Bar, R., S. Itzkovitz, et al. (2007). "Functional atlas of the integrin adhesome." Nat Cell Biol **9**(8): 858-867.

Zaiss, D. M., N. de Graaf, et al. (2008). "The proteasome immunosubunit multicatalytic endopeptidase complex-like 1 is a T-cell-intrinsic factor influencing homeostatic expansion." Infect Immun **76**(3): 1207-1213.

Zhang, C., D. R. Dowd, et al. (2003). "Nuclear coactivator-62 kDa/Ski-interacting protein is a nuclear matrix-associated coactivator that may couple vitamin D receptor-mediated transcription and RNA splicing." J Biol Chem **278**(37): 35325-35336.

Zhang, J., N. Naslavsky, et al. (2012). "Rabs and EHDs: alternate modes for traffic control." Biosci Rep **32**(1): 17-23.

8 Appendix 1 - PlateletWeb user guide

(This information is also published on Blood journal and can be found at Boyanova et. al (*Boyanova, Nilla et al. 2012*)).

1. Introduction

1.1 Using the PlateletWeb knowledge base

PlateletWeb is a bioinformatical knowledge base covering the platelet proteome, transcriptome and interactome. By combining data of published platelet proteome and transcriptome studies with comprehensive protein-protein interaction data a first model of the platelet interactome has been derived. *PlateletWeb* makes this platelet interactome accessible to all researchers and allows an easy navigation through the web of platelet interactions. Additional information about the type of evidence (proteome, transcriptome or both) for each protein is provided. Based on the data of a recent phosphoproteome study the type of phosphorylation (serine/threonine, tyrosine) is indicated by an icon on the right side.

1.2 Search for a specific protein

Searching for a specific protein is easy: Simply enter the name of the protein or a part of the name (at least three characters) into the search field and submit your query. You will obtain a list of all proteins matching your query term. If the query term is too general it will produce too many (>50) matches. In this case no results will be returned and you will be asked to specify your search.

1.3 Navigate through the platelet interactome and phosphoproteome

The result list contains the name and a short description for each protein. The marker on the left side indicates whether the protein has been described on the level of the proteome, transcriptome or both. On the right hand side information about the phosphorylation status of each protein in unstimulated platelets is provided. Most importantly the link beneath the protein description displays the number of interaction partners in the platelet. By clicking on this link you will obtain the list of interaction partners (all interactants or only platelet interactants). This allows you to navigate through the entire interactome network.

2. Overview / Legend

2.1 Platelet Proteins

Icons describing the level of detection for each protein

Proteome	**Proteome** proteins found on the level of the platelet proteome
SAGE	**SAGE** proteins found on the level of the platelet transcriptome
Both	**Both** proteins found on the level of the platelet proteome as well as on the level of the transcriptome
None	**None** proteins which have not been detected in the platelet

2.2 Type of Experiment

Icons describing the type of experiment which detects the interactions and phosphorylations of the proteins

⬢	*in vivo* Experiment
⬢	*in vitro* Experiment
⬢	Yeast-two-hybrid (Y2H) Experiment

2.3 Protein Phosphorylations

Icons describing the type of phosphorylations of each protein

ST	**Experimental Platelet Phosphorylation S/T** denotes platelet proteins phosphorylated on *serine or/and threonine* residues. The phosphosites are detected in platelets
Y	**Experimental Platelet Phosphorylation Y** denotes platelet proteins phosphorylated on a *tyrosine* residue. The phosphosites are detected in **platelets**
ST Y	**Experimental Platelet Phosphorylation S/T/Y** denotes proteins phosphorylated on *serine or/and threonine* as well as *tyrosine* residues. The phosphosites are detected in **platelets**
ST	**Experimental Human Phosphorylation S/T** denotes proteins phosphorylated on *serine or/and threonine* residues. The phosphosites are detected in **human cells**
Y	**Experimental Human Phosphorylation Y** denotes proteins phosphorylated on a *tyrosine* residue. The phosphosites are detected in **human cells**
ST Y	**Experimental Human Phosphorylation S/T/Y** denotes proteins phosphorylated on *serine or/and threonine* as well as *tyrosine* residues. The phosphosites are detected in **human cells**

2.4 Kinase Type

Icons describing the type of kinase

S/T	**S/T Kinase** denotes kinases that phosphorylate targets on *serine or/and threonine* residues
Y	**Y Kinase** denotes kinases that phosphorylate targets on *tyrosine* residues

dual	**S/T/Y Kinase** denotes kinases that phosphorylate targets on *serine or/and threonine* as well as *tyrosine* residues

2.5 Phosphatases

Icons describing the Phosphatase

P	**Phosphatase** denotes manually curated protein phosphotases

2.6 Sources of Information

Icons describing the source of information about Interactions and phosphorylations of the proteins

🟠	**HPRD** denotes HPRD (Human Protein Reference Database) as the source of phosphorylation information
🔵	**PhosphoSite** denotes PhosphoSite as the source of phosphorylation information
🟡	**NetworKIN** denotes NetworKIN as the source of kinase predictions for experimental platelet phosphorylations

2.7 Interaction Type

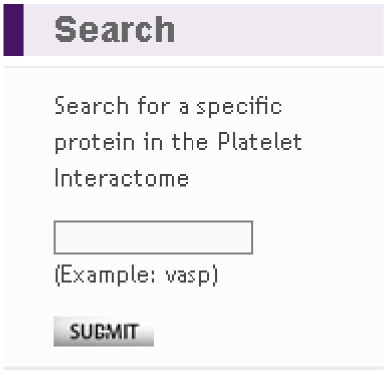
Icons describing the Interaction type of the proteins

⇨	**Protein-protein interaction** denotes interactions between proteins based on the NCBI dataset

	(Entrez gene)
➡	**Phosphorylation** denotes phosphorylation reactions (kinases) derived from annotations in the HPRD and PhosphoSite databases
➡	**Kinase predictions** denotes phosphorylation reactions for kinases predicted by NetworKIN for experimental platelet phosphorylations
➡	**Dephosphorylation** denotes dephosphorylation reactions (phosphatases) derived from annotations in the HPRD database

3. Where do I begin?

PlateletWeb is a comprehensive systems biology tool, which allows a thorough analysis of HUMAN platelet and non-platelet proteins, their interactions, phosphorylation state and physical characteristics. The database contains validated experimental data on the platelet proteome and phosphoproteome, along with literature-derived information from HPRD (Human Protein Database Version 9.0). It is the first database of its kind, which gives the opportunity for extracting interactive subnetworks of proteins on a platelet and non-platelet level. The user can choose between regular search of a specific protein of interest OR advanced search providing detailed information.



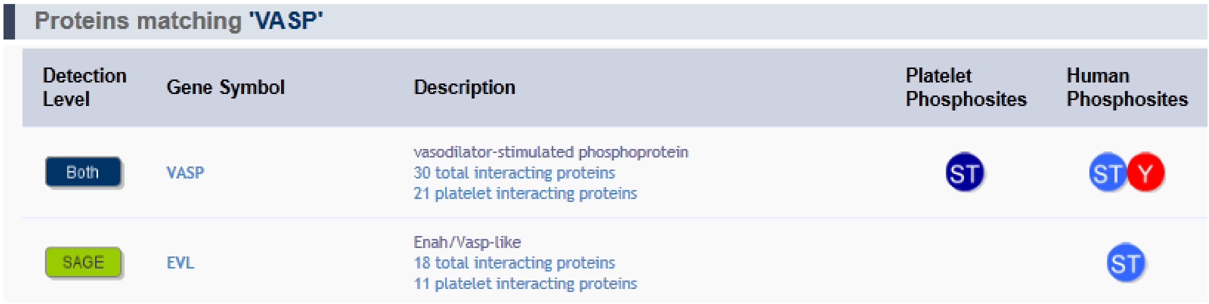
Search for a specific protein in the Platelet Interactome

(Example: vasp)

Search

In the search field on the left side of the window, just type in the name of the protein you are interested in and click on the "SUBMIT" button in order to get the protein related information.

This retrieves the list of proteins matching your search term. In our example, our search term is "VASP" and so we got the proteins matching VASP as output.

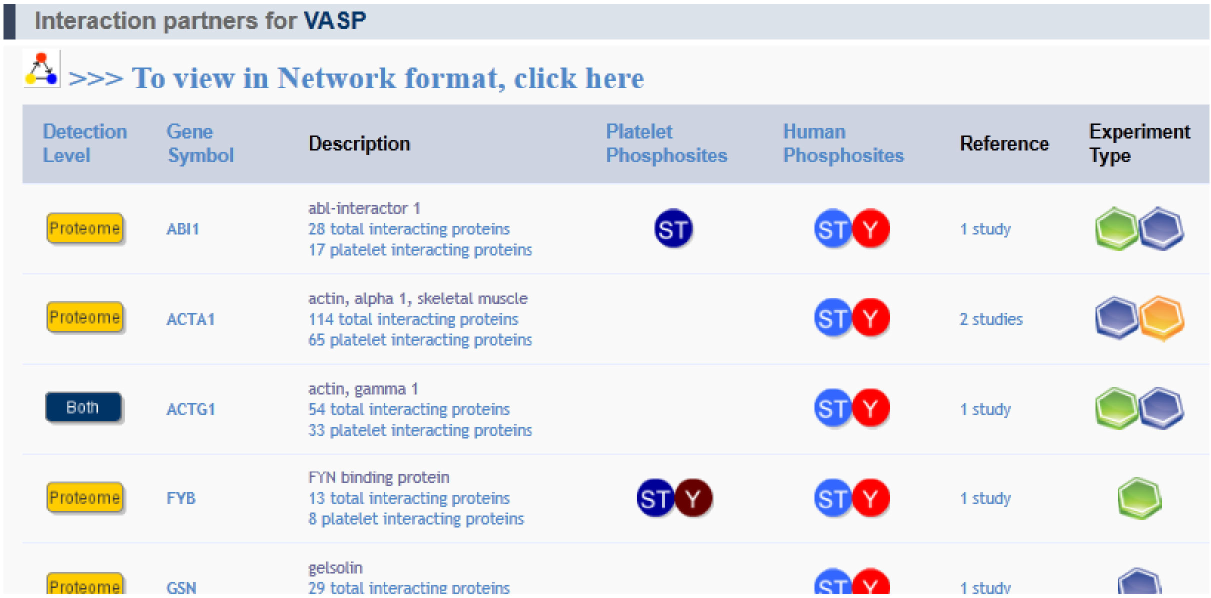


Proteins matching 'VASP'

Detection Level	Gene Symbol	Description	Platelet Phosphosites	Human Phosphosites
Both	VASP	vasodilator-stimulated phosphoprotein 30 total interacting proteins 21 platelet interacting proteins	ST	ST Y
SAGE	EVL	Enah/Vasp-like 18 total interacting proteins 11 platelet interacting proteins		ST

Click on "platelet interacting proteins" to view the list of platelet interactions (In the picture, a subset of the proteins are shown).

Alternatively, you can also click on "Total interacting proteins" for the complete list of interactions with this protein.



Interaction partners for VASP

>>> To view in Network format, click here

Detection Level	Gene Symbol	Description	Platelet Phosphosites	Human Phosphosites	Reference	Experiment Type
Proteome	ABI1	abl-interactor 1 28 total interacting proteins 17 platelet interacting proteins	ST	ST Y	1 study	⬢ ⬢
Proteome	ACTA1	actin, alpha 1, skeletal muscle 114 total interacting proteins 65 platelet interacting proteins		ST Y	2 studies	⬢ ⬢
Both	ACTG1	actin, gamma 1 54 total interacting proteins 33 platelet interacting proteins		ST Y	1 study	⬢ ⬢
Proteome	FYB	FYN binding protein 13 total interacting proteins 8 platelet interacting proteins	ST Y	ST Y	1 study	⬢
Proteome	GSN	gelsolin 29 total interacting proteins		ST Y	1 study	⬢

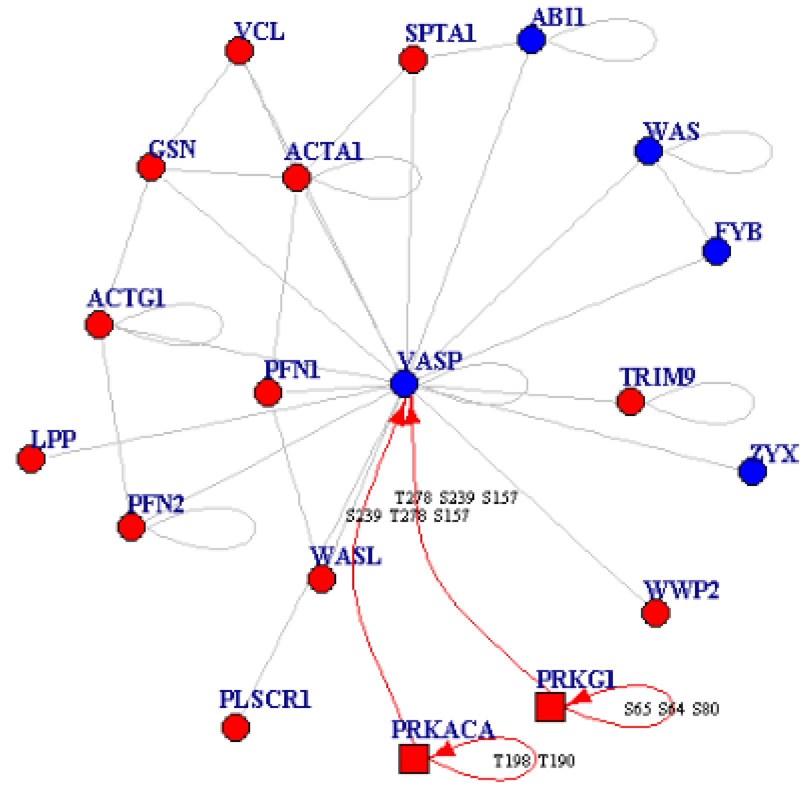
To view in picture format:

On the top of the interactions page, you will find the link "To View in Network format, click here" which leads to the graphical representation of all retrieved interactions with the searched protein.

What Next: Using the links "Please click here to download the file in pdf format" and "Click here to download the compressed file for Cytoscape" you can either view and save the image in the pdf (Adobe Reader) format or alternatively download the information to view in Cytoscape.

Finally,

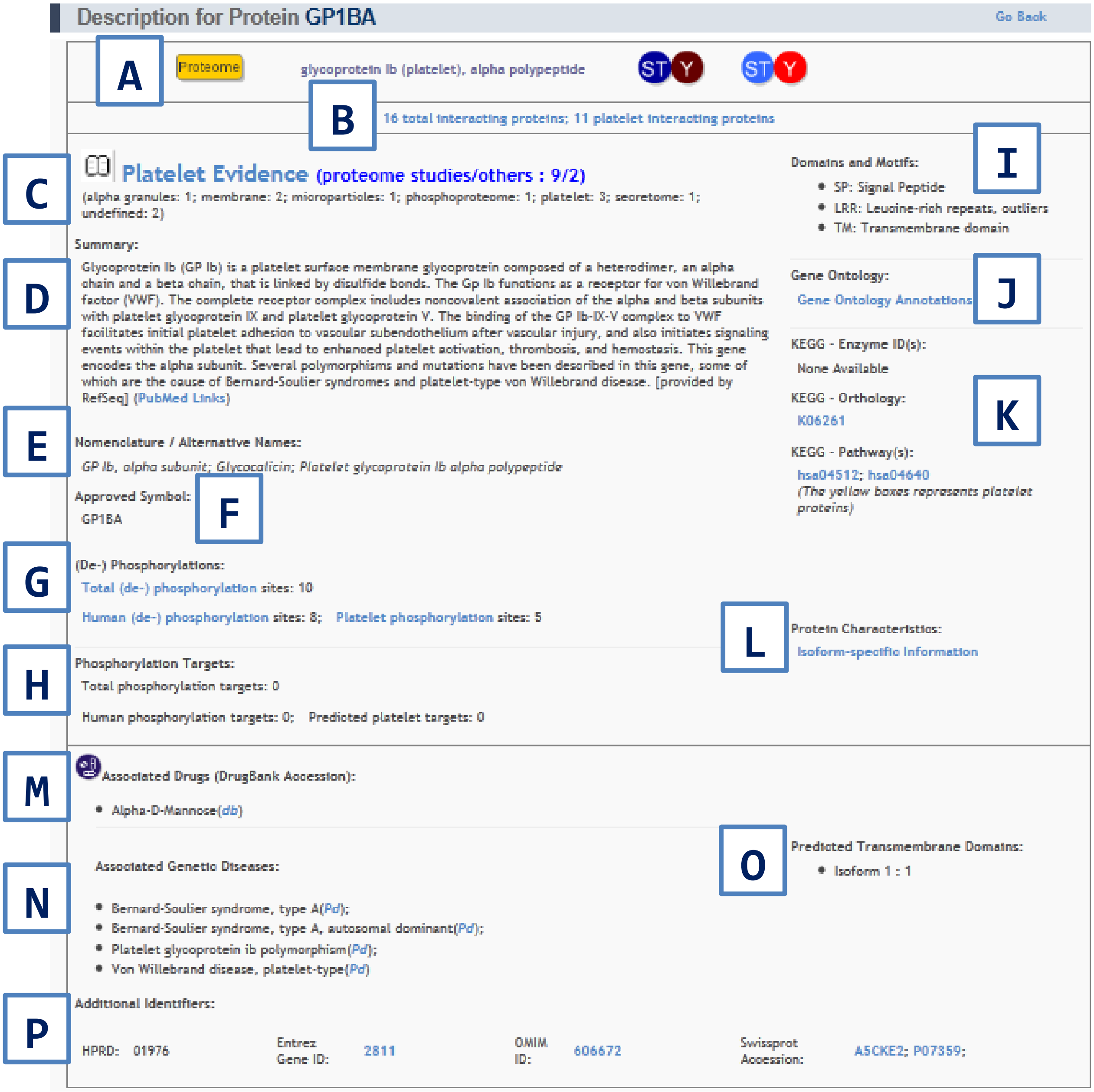
As we can see, just by providing the name of the protein, one can easily navigate through all its interactions, get the complete description, view and/or download the network graph associated with it. Apart from this, there is always the possibility of searching for proteins with their physical characteristics, (using advanced search), download the subnetwork for a specific set of proteins (Cytoscape Download), and retrieve the complete information about their phosphorylation state, and if the proteins are kinases, their phosphorylation targets.



4. Description of Protein

Once the protein's name is entered in the search field, a list of proteins matching the search term appears. By clicking on the protein's name in the list, the user can then open the page containing its description. This page includes information about the protein’s approved symbol, its multiple identifiers, the study and the fraction where the protein has been detected, its phosphorylation sites and alternative names. It also contains additional protein characteristics such as protein domains and motifs, Gene Ontology (GO) terms and predictions for transmembrane domains along with drugs and genetic diseases associated with the protein.

In the special case of kinases, it is possible to examine all phosphorylation targets of a specific kinase and link to the targets' description directly.



Description for Protein GP1BA

glycoprotein Ib (platelet), alpha polypeptide

16 total interacting proteins; 11 platelet interacting proteins

Platelet Evidence (proteome studies/others : 9/2)
(alpha granules: 1; membrane: 2; microparticles: 1; phosphoproteome: 1; platelet: 3; secretome: 1; undefined: 2)

Summary:
Glycoprotein Ib (GP Ib) is a platelet surface membrane glycoprotein composed of a heterodimer, an alpha chain and a beta chain, that is linked by disulfide bonds. The Gp Ib functions as a receptor for von Willebrand factor (VWF). The complete receptor complex includes noncovalent association of the alpha and beta subunits with platelet glycoprotein IX and platelet glycoprotein V. The binding of the GP Ib-IX-V complex to VWF facilitates initial platelet adhesion to vascular subendothelium after vascular injury, and also initiates signaling events within the platelet that lead to enhanced platelet activation, thrombosis, and hemostasis. This gene encodes the alpha subunit. Several polymorphisms and mutations have been described in this gene, some of which are the cause of Bernard-Soulier syndromes and platelet-type von Willebrand disease. [provided by RefSeq] (PubMed Links)

Nomenclature / Alternative Names:
GP Ib, alpha subunit; Glycocalicin; Platelet glycoprotein Ib alpha polypeptide

Approved Symbol:
GP1BA

(De-) Phosphorylations:
Total (de-) phosphorylation sites: 10
Human (de-) phosphorylation sites: 8; Platelet phosphorylation sites: 5

Phosphorylation Targets:
Total phosphorylation targets: 0
Human phosphorylation targets: 0; Predicted platelet targets: 0

Domains and Motifs:
- SP: Signal Peptide
- LRR: Leucine-rich repeats, outliers
- TM: Transmembrane domain

Gene Ontology:
Gene Ontology Annotations

KEGG - Enzyme ID(s):
None Available

KEGG - Orthology:
K06261

KEGG - Pathway(s):
hsa04512; hsa04640
(The yellow boxes represents platelet proteins)

Protein Characteristics:
Isoform-specific Information

Associated Drugs (DrugBank Accession):
- Alpha-D-Mannose(*db*)

Associated Genetic Diseases:
- Bernard-Soulier syndrome, type A(*Pd*);
- Bernard-Soulier syndrome, type A, autosomal dominant(*Pd*);
- Platelet glycoprotein ib polymorphism(*Pd*);
- Von Willebrand disease, platelet-type(*Pd*)

Predicted Transmembrane Domains:
- Isoform 1 : 1

Additional Identifiers:

HPRD:	01976	Entrez Gene ID:	2811	OMIM ID:	606672	Swissprot Accession:	A5CKE2; P07359;

A. General information about the protein

The protein detection level along with its phosphorylated residues is presented.

B. Total Interacting Proteins / Platelet Interacting Proteins

The number of human and platelet interacting partners with links to the interacting proteins and network view.

C. *Platelet Evidence*

Clicking on this link would provide a list of proteome studies, in which the protein was detected. Furthermore, it contains information about the cellular compartment of the protein along with a link to the study. The studies are divided into 2 groups, proteome data and database/SAGE source information. Additionally, the studies are categorized according to the cell fraction of the protein and the number of appearances in each fraction is displayed.

D. *Summary*

A short description of the protein and its functions is presented along with the links to PubMed.

E. *Nomenclature / Alternative Names*

Other names assigned to the protein.

F. *Approved Symbol*

This is the HUGO approved Symbol of the protein.

G. *(De-) Phosphorylations*

Total phosphorylations, divided into human phosphorylation sites and platelet phosphorylation sites.

(see Section 5 for additional information)

H. *Phosphorylation Targets*

Kinases and Phosphatases are provided with a list of all their platelet and human substrates.

I. Domains and Motifs

All domains and motifs of the protein are listed with a short abbreviation of the domain name and a complete domain description where available.

J. Gene Ontology

Full listing of all GO annotations of the protein. Further accessible are the GO terms, which contain at least 1 but not more than 150 platelet proteins. A link to all related children terms is displayed next to each of the protein's GO terms.

K. KEGG Information

Information about KEGG enzyme classification (for enzymes), orthology along with graphical presentation of pathways containing the platelet proteins (in yellow).

L. Protein Physical Characteristics

Isoform-specific information about the physical properties of the protein. Further details on the protein's molecular weight, isoform accession, its length, isoelectric point and the protein sequence in a downloadable FASTA format can be accessed.

M. Associated Drugs (DrugBank Associations)

All drugs are listed with their corresponding accession number for DrugBank. Clicking on the link "(db)" retrieves the DrugBank webpage with detailed information about the drug.

N. Associated Genetic Diseases

The "(Pd)" link next to each genetic disease term navigates to the PubMed information source.

O. Predicted Transmembrane Domains

Information on the number of predicted transmembrane domains for each isoform of the protein is shown here.

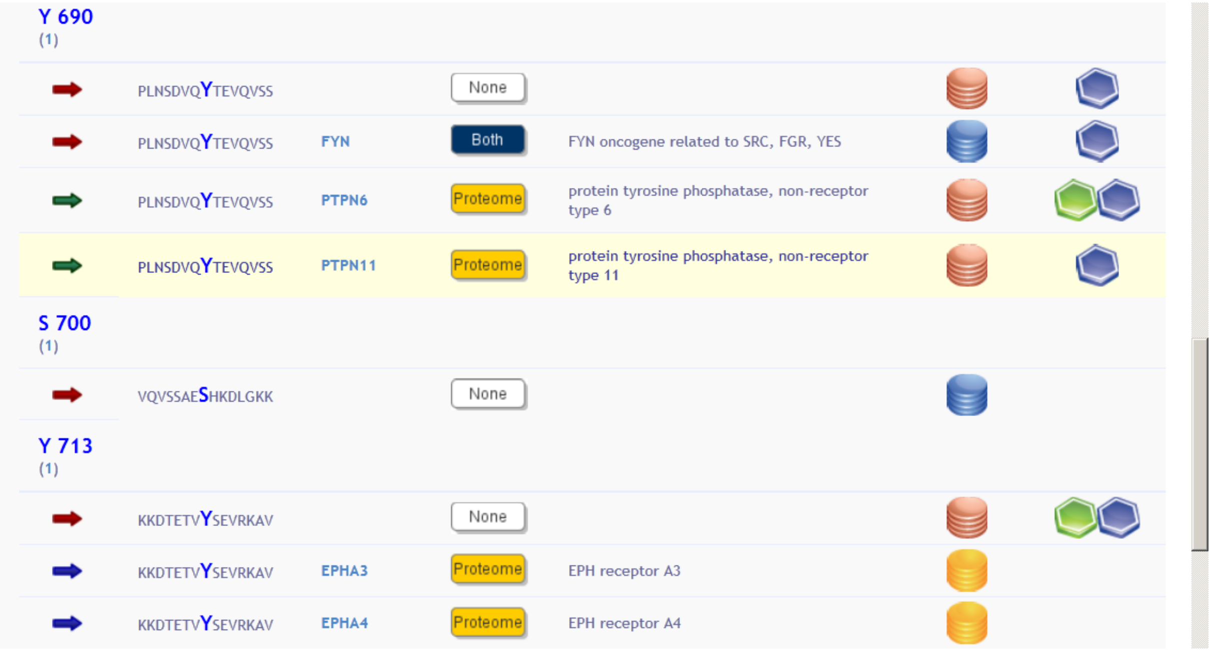
P. Additional Identifiers

External links associated to the protein - identifiers from HPRD, NCBI gene id, OMIM id and the Swissprot accession number are provided here.

5. Phosphorylation and Kinase information

Query: PECAM-1 phosphorylations (example)

Result:



Y 690 (1)						
➡	PLNSDVQ**Y**TEVQVSS		None		●	⬢
➡	PLNSDVQ**Y**TEVQVSS	FYN	Both	FYN oncogene related to SRC, FGR, YES	●	⬢
➡	PLNSDVQ**Y**TEVQVSS	PTPN6	Proteome	protein tyrosine phosphatase, non-receptor type 6	●	⬢⬢
➡	PLNSDVQ**Y**TEVQVSS	PTPN11	Proteome	protein tyrosine phosphatase, non-receptor type 11	●	⬢
S 700 (1)						
➡	VQVSSAE**S**HKDLGKK		None		●	
Y 713 (1)						
➡	KKDTETV**Y**SEVRKAV		None		●	⬢⬢
➡	KKDTETV**Y**SEVRKAV	EPHA3	Proteome	EPH receptor A3	●	
➡	KKDTETV**Y**SEVRKAV	EPHA4	Proteome	EPH receptor A4	●	

Clicking on the "(De-) Phosphorylations" link retrieves a list of all phosphorylated and de-phosphorylated residues and their position in the protein's sequence. It is also possible to view the human (de-) phosphorylations and platelet phosphorylations individually.

Underneath this information, you can find the isoforms in which the site is identified. Further down within the same phosphosite, one can find the phosphorylated motif sequence with the site highlighted in the middle. Additionally, the phosphorylating kinase (if available) is shown next to it. The source of phosphorylation information is presented under Reference. For kinase predictions on platelet phosphorylation sites the reference is given as NetworKIN (prediction algorithm). For all data from literature, the reference is either PhosphoSite or HPRD (Source: PubMed). Finally, each phosphorylation is presented with the type of experiment used for its detection. For some of the experimentally validated phosphorylations, there is a NetworKIN kinase prediction displayed on the description page of the protein. The HPRD-derived kinase-substrate pairs are also listed, but are not necessarily proven to play a role in platelets specifically. Still, it is a strong indication if both the kinase and its target protein are found in platelets and the target is phosphorylated.

6. Advanced Search Features

The advanced search features allow combining different search options into a more complex search with an emphasis on platelet proteins to obtain more specific results.

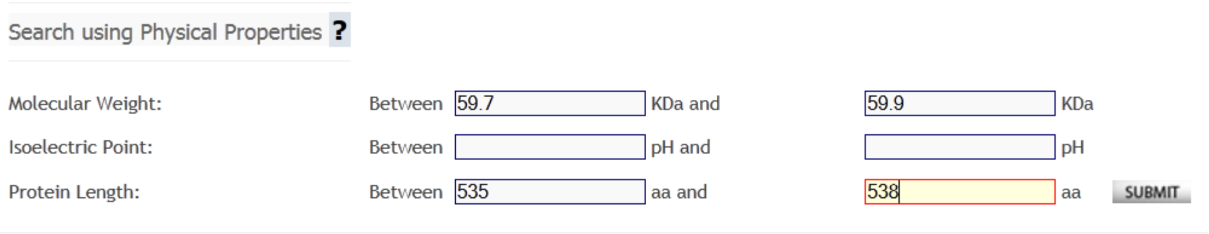
6.1 Keyword search

As an example, the search term “**rheumatoid arthritis**” reveals 27 total human proteins of which 10 are identified in platelets.

6.2 Physical property search

Search for proteins in a particular range according to their molecular weight, isoelectric point or protein length.

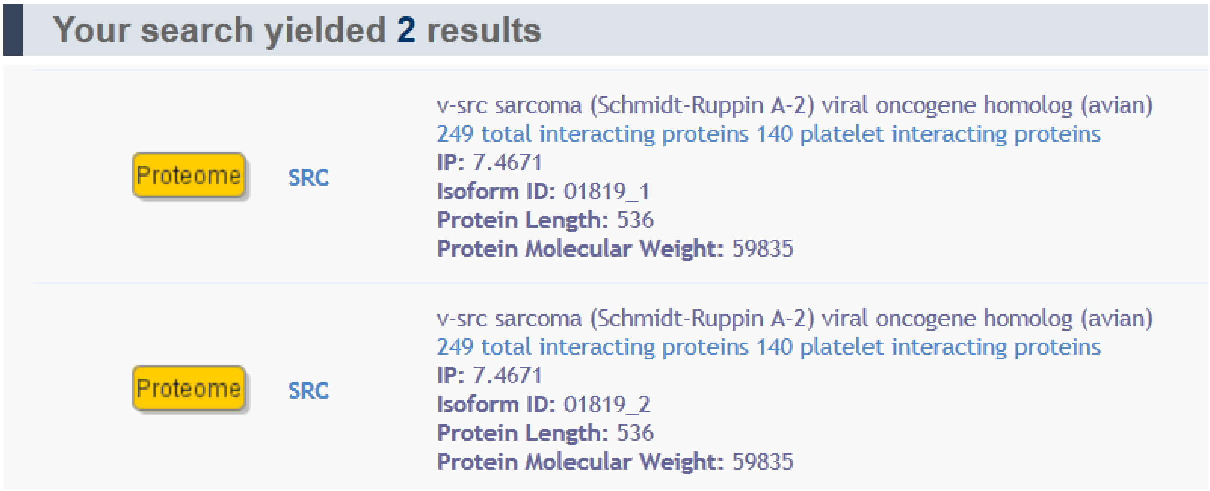
Query:



Search using Physical Properties ?

Molecular Weight:	Between	59.7	KDa and	59.9	KDa
Isoelectric Point:	Between		pH and		pH
Protein Length:	Between	535	aa and	538	aa

Result:



Your search yielded 2 results

v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
249 total interacting proteins 140 platelet interacting proteins
IP: 7.4671
Isoform ID: 01819_1
Protein Length: 536
Protein Molecular Weight: 59835

v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
249 total interacting proteins 140 platelet interacting proteins
IP: 7.4671
Isoform ID: 01819_2
Protein Length: 536
Protein Molecular Weight: 59835

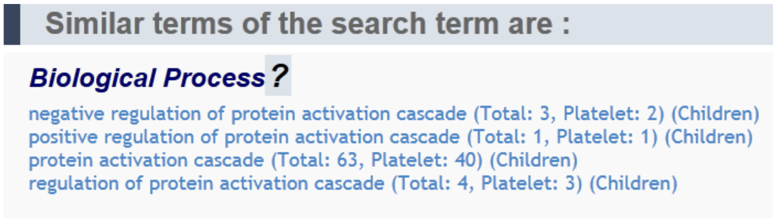
Note: Both results are SRC, however they belong to two different isoforms. It is possible to further navigate to the protein description page and the interactions and phosphorylations page from here.

6.3 Gene ontology Search

Query:

Search for the term “protein activation”

Result:



Similar terms of the search term are :

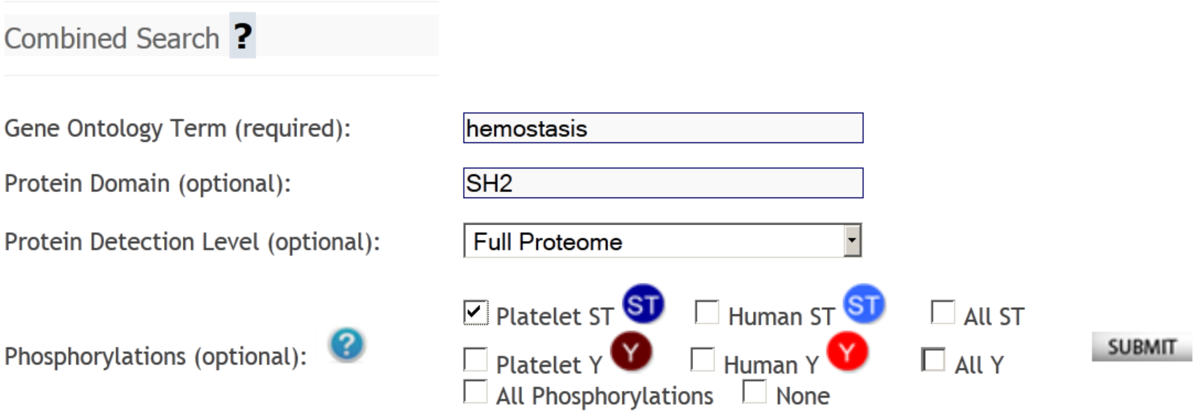
Biological Process ?

negative regulation of protein activation cascade (Total: 3, Platelet: 2) (Children)
positive regulation of protein activation cascade (Total: 1, Platelet: 1) (Children)
protein activation cascade (Total: 63, Platelet: 40) (Children)
regulation of protein activation cascade (Total: 4, Platelet: 3) (Children)

6.4 Combination search

A combination of functional enrichment, protein domains and the detection level along with the phosphorylation information. This search option is very useful for a given biological question, for example connected to platelet activatory (or inhibitory) processes.

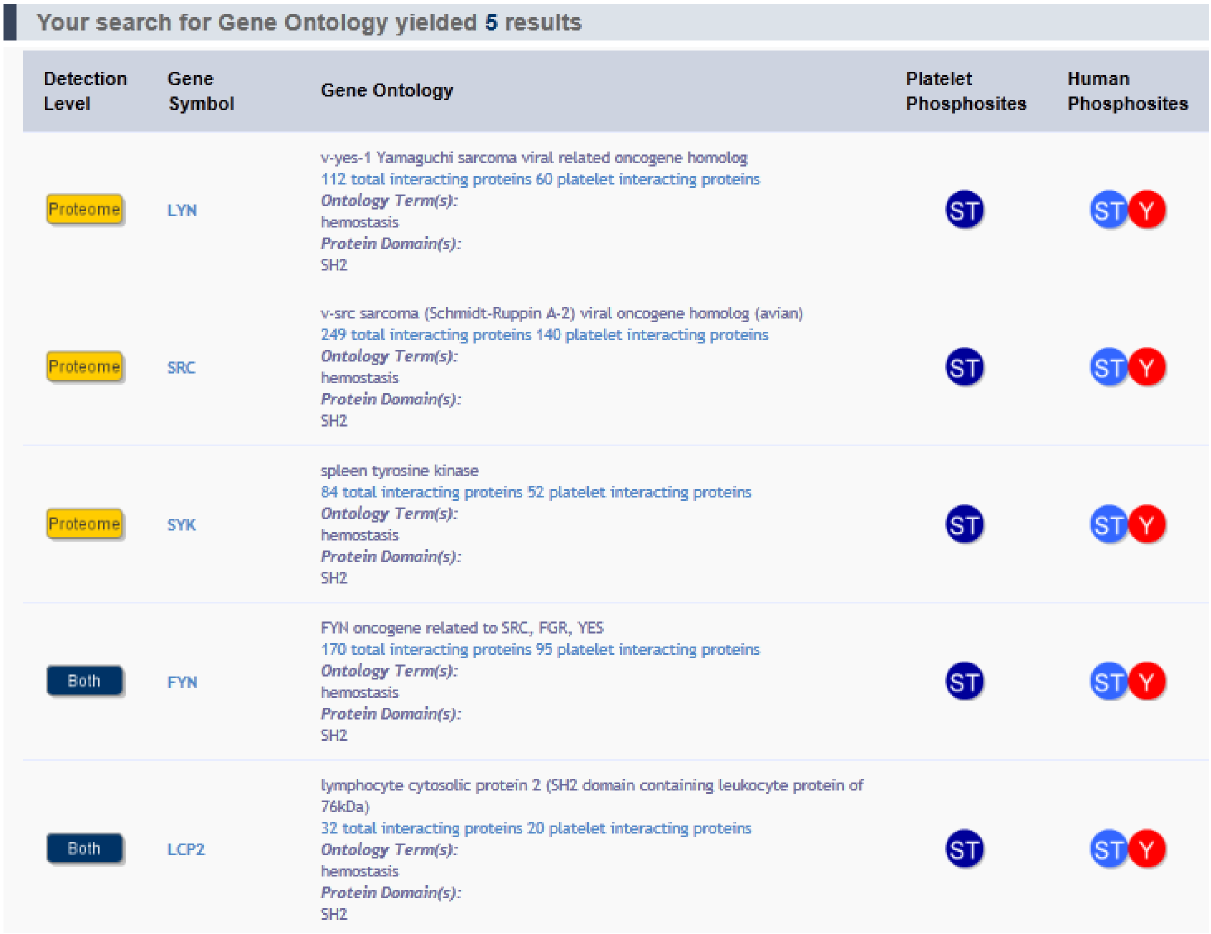
Query: As an example, a search can be performed identifying hemostasic proteins, which bind tyrosine phosphorylated residues and therefore contain an SH2 domain and are additionally phosphorylated on a serine or a threonine residue, detected in a platelet proteome experiment.



Combined Search ?

Gene Ontology Term (required):	hemostasis
Protein Domain (optional):	SH2
Protein Detection Level (optional):	Full Proteome
Phosphorylations (optional):	☑ Platelet ST □ Human ST □ All ST □ Platelet Y □ Human Y □ All Y □ All Phosphorylations □ None

Result: This search results in 5 platelet proteins having the given characteristics.



Your search for Gene Ontology yielded 5 results

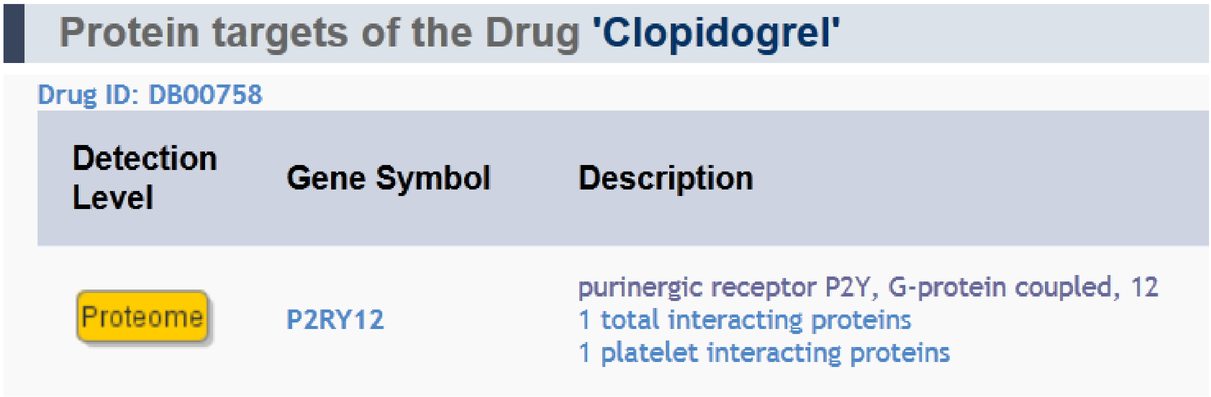
Detection Level	Gene Symbol	Gene Ontology	Platelet Phosphosites	Human Phosphosites
Proteome	LYN	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog 112 total interacting proteins 60 platelet interacting proteins *Ontology Term(s):* hemostasis *Protein Domain(s):* SH2	ST	ST Y
Proteome	SRC	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian) 249 total interacting proteins 140 platelet interacting proteins *Ontology Term(s):* hemostasis *Protein Domain(s):* SH2	ST	ST Y
Proteome	SYK	spleen tyrosine kinase 84 total interacting proteins 52 platelet interacting proteins *Ontology Term(s):* hemostasis *Protein Domain(s):* SH2	ST	ST Y
Both	FYN	FYN oncogene related to SRC, FGR, YES 170 total interacting proteins 95 platelet interacting proteins *Ontology Term(s):* hemostasis *Protein Domain(s):* SH2	ST	ST Y
Both	LCP2	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa) 32 total interacting proteins 20 platelet interacting proteins *Ontology Term(s):* hemostasis *Protein Domain(s):* SH2	ST	ST Y

6.5 Drug search

Query:

Search the drug name “clopidogrel” for its drug targets reveals,

Result:



Protein targets of the Drug 'Clopidogrel'

Drug ID: DB00758

Detection Level	Gene Symbol	Description
Proteome	P2RY12	purinergic receptor P2Y, G-protein coupled, 12 1 total interacting proteins 1 platelet interacting proteins

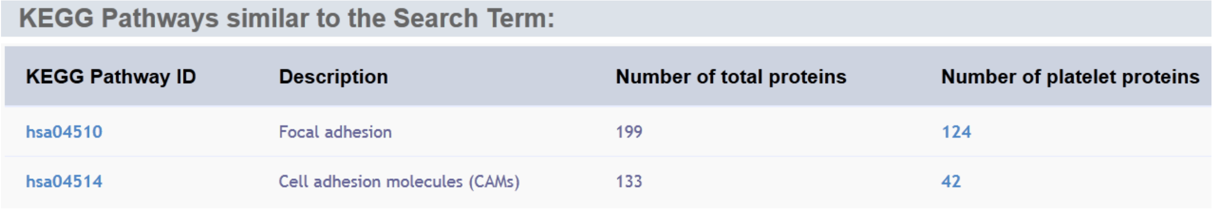
The DrugBank source page can be accessed from the link in the resulting panel (Drug ID: DB00758).

6.6 Pathway search

Query:

Search for the similar terms of “adhesion” in KEGG pathway reveals

Result:



KEGG Pathways similar to the Search Term:

KEGG Pathway ID	Description	Number of total proteins	Number of platelet proteins
hsa04510	Focal adhesion	199	124
hsa04514	Cell adhesion molecules (CAMs)	133	42

6.7 Subnetwork extraction

The "Extract subnetwork" option in the advanced search is introduced for the analysis of a set of proteins. The aim of this feature is to extract networks of interest from *PlateletWeb* and to visualize the interactions in a group of pre-selected proteins. The names of the proteins are entered in the search field, separated by a tab or comma. Click on the submit button to go to the next step.

The resulting list of proteins is supplemented by a list of similar names in case proteins were not found in the search. Non-platelet proteins can also be added and analyzed. If the proper list of proteins is already available, the user can now choose between a graphical presentation of the constructed network and a list of interactions given in a text-format.

By clicking on the "Network format" the network appears on the screen, depicting the proteins as nodes and the interactions between them as edges. All interactions and phosphorylation states are color and shape coded in the final result.

Additionally, an option for network download is available by saving the network in a pdf file.

If the user prefers extracting the data and creating a visualization with Cytoscape, there is a compressed Cytoscape file available for download. It contains a .sif file of the extracted interactome network, as well as edge and node attributes files (.eda, .noa), which can be added into Cytoscape to visualize specific features of the network.

The edge attribute file contains the phosphorylation sites for each phosphorylated protein and the node attribute files include information about the node itself: name of the protein, whether the protein is a "kinase or not", whether it is "phosphorylated or not" etc. An easy option is to save the file on the desktop and then extract the folders. The visualization characteristics and legends are contained in the .props file in the folder "properties".

Steps for importing the files into Cytoscape:

File -> Import -> Network (multiple file types) -> .sif file from tmp folder

File -> Import -> Node attributes -> .noa files from tmp folder

File -> Import -> Edge attributes -> .eda files from tmp folder

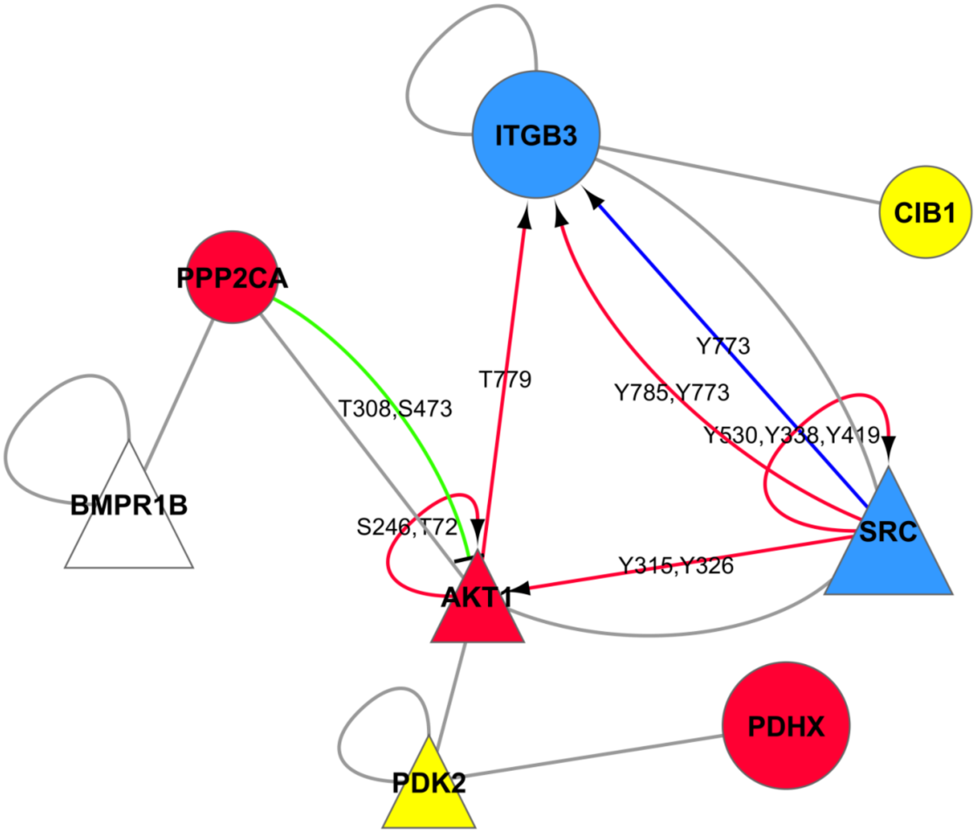
File -> Import -> Vizmap property file -> cytoscapevisuals.props from properties folder

In order to view the optimal network in Cytoscape, select Layout -> yFiles -> Organic. The final version of the created network should be visible and available for further analysis.

Query:

ITGB3, SRC, CIB1, AKT1, PDK2, PPP2CA, BMPR1B, PDHX

Result:


ITGB3
CIB1
PPP2CA
BMPR1B
T308,S473
T779
Y773
Y785,Y773
Y530,Y338,Y419
S246,T72
SRC
AKT1
Y315,Y326
PDHX
PDK2

9 Table of Figures

10 Table of Tables

11 List of publications

Boyanova, D*., S. Nilla*, et al. (2012). "PlateletWeb: a systems biologic analysis of signaling networks in human platelets." Blood **119**(3): e22-34. (Equal contributions)

Nilla S*., Boyanova D*., et al. (2012). "Identifying functional modules in protein-protein interaction networks using exact solutions and semantic similarity." (Ready for submission), (Equal contributions)

12 Conference contributions

GTH 2010, Nürnberg

D. Boyanova, **S. Nilla**, I. Birschmann, U. Walter, T. Dandekar, M. Dittrich
"*PlateletWeb*: An integrated Systems Biology platform for the analysis of Platelet Signaling"
1st Joint Meeting GTH & NVTH, Nürnberg (February 24th - 27th), 2010

SBMC 2010, Freiburg

D. Boyanova, **S. Nilla**, I. Birschmann, U. Walter, T. Dandekar, M. Dittrich
"Unravelling Cellular Networks: A systems biological perspective on platelet signaling"
3rd Conference on Systems Biology of Mammalian Cells, Freiburg (June 3rd - 5th), 2010

G. Wangorsch, M. Mischnik, K. Glausauer, **S. Nilla**, D. Boyanova, A. Sickmann, M. Dittrich, J. Timmer, J. Geiger, T. Dandekar
"PGI2 and ADP P2Y12 receptor signaling: downstream events and crosstalk"
3rd Conference on Systems Biology of Mammalian Cells, Freiburg (June 3rd - 5th), 2010

Boston 2010

G. Wangorsch, D. Boyanova, **S. Nilla**, M. Dittrich, T. Dandekar
"Integrating platelet proteome, phosphoproteome and drug information for a systems biological analysis of pharmacological targets"
International Conference on Systems Biology of Human Disease 2010 in Boston, MA, USA (June 16th -18th), 2010

GTH 2011, Wiesbaden

D. Boyanova, **S. Nilla**, I. Birschmann, U. Walter, T. Dandekar, M. Dittrich
"Integrated data analysis of functional protein networks in human platelets"
55th Annual Meeting of the GTH, Wiesbaden (February 16th-19th), 2011

G. Wangorsch, **S. Nilla**, D. Boyanova, M. Dittrich, T. Dandekar
"Probing modulatory interactions of the vWF/GP1b signaling cascade"
55th Annual Meeting of the GTH, Wiesbaden (February 16th-19th), 2011

M. Dittrich, D. Boyanova, **S. Nilla**, M. Balz, T. Dandekar, I. Birschmann
"An Integrated network of hemostatic drugs and drug targets in human platelets"
55th Annual Meeting of the GTH, Wiesbaden (February 16th-19th), 2011

ISMB 2011, Vienna

D.Boyanova, **S. Nilla**, D. Beisser, G. Klau, T. Dandekar, T. Müller, M. Dittrich
"Integrated analysis of cellular networks using phosphoproteome data to identify active signaling modules"

S. Nilla, D. Boyanova, D. Beisser, G. Klau, T. Dandekar, T. Müller, M. Dittrich
"Identifying functional modules in protein-protein interaction networks based on semantic similarity using an exact approach"
19th Annual International Conference on Intelligent Systems for Molecular Biology, Vienna (June 16th – 19th), 2011

European Conference

A. Zeeshan, M. Saman, C. Liang, A. Cecil, G. Wangorsch, **S. Nilla**, D. Boyanova, M. Naseem, A. Fieselmann, M. Dittrich., T. Dandekar

"Intelligent Information Management for efficient computational biology"

ICT 2011 - Information and Networking Day - Intelligent Information Management Jean Monnet Conference Centre, Luxembourg (26 September 2011)

Regensburg 2011

E. Schneider, A. Keller, **S. Nilla**, L. R. Jensen, I. Kondova, R. Farcas, E. Fuchs, A. Kuss, T. Haaf and T. Dandekar

"HAR4 and human brain development - new insights from comparative gene expression, 2D-3D structures and Interactome"

German Society of Human Genetics, Regensburg, 2011

Freising 2011

Cecil A, **Nilla S**, Schaefer B, Sotriffer C, Dandekar T

"DrugPoint – a retrieval software and databank to connect proteins, drugs and targets"

GCB 2011 - German Conference on Bioinformatics 2011

Acknowledgements

Many individuals have supported me during the development of this thesis to whom I would like to express my deepest gratitude.

First and foremost, I would like to address my gratitude to *Prof. Dr. Thomas Dandekar* for giving me the possibility to do this PhD-thesis at the Department of Bioinformatics and for supporting me during the work. I would like to extend my gratitude to *Prof. Christian Wegener* for his time and willingness to serve on my thesis committee.

I am grateful to *Dr. Dr. Marcus Dittrich* and *Dr. Tobias Müller* for their guidance and motivation during my thesis and I thank them both for their time, support and most of all, the encouragement they provided me throughout my project. They strongly contributed to my ambitious ideas and overall success of this project.

My cordial thanks are also addressed to *Prof. Dr. Jörg Schultz* along with the group members of Bioinformatics department for their contributions in the meetings and for their fruitful discussions and the ideas which helped to give my best to the world of Science. A special recognition for *Gaby Wangorsch*, *Christian Koetschan, Daniela Beißer* and *Alexander Cecil* has to be given as they were always there for me in every aspect with their incredible help and with ever ending support. A special mention is to *Desislava Boyanova*, who supported me in every possible way, keep up to the pace in unlimited discussions, tune in new ideas, and defines the true meaning of the word Team work. Her contribution into my thesis was impeccable. True friends are rare, but these are just the best.

A special recognition is to *Daniel Laible*, *Carmen Voit* and *Tobias Hartmann* for their never ending support and not to forget, their special attempt to teach me good German.

Last but not least, I would like to thank my parents, my brother and my wife for their unlimited support in my life.