INAUGURAL-DISSERTATION

zur Erlangung der Doktorwürde der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Apichat Suratanee, M.Sc.
aus Bangkok, Thailand

Tag der mündlichen Prüfung: 15.10.2012

Computational Analysis of RNAi Screening Data to Identify Host Factors Involved in Viral Infection and to Characterize Protein-Protein Interactions

Gutachter: Prof. Dr. Roland Eils
Prof. Dr. Gerhard Reinelt

Abstract

The study of gene functions in a variety of different treatments, cell lines and organisms has been facilitated by RNA interference (RNAi) technology that tracks the phenotype of cells after silencing of particular genes. In this thesis, I describe two computational approaches developed to analyze the image data from two different RNAi screens. Firstly, I developed an alternative approach to detect host factors (human proteins) that support virus growth and replication of cells infected with the Hepatitis C virus (HCV). To identify the human proteins that are crucial for the efficiency of viral infection, several RNAi experiments of viral-infected cells have been conducted. However, the target lists from different laboratories have shown only little overlap. This inconsistency might be caused not only by experimental discrepancies, but also by not fully explored possibilities of the data analysis. Observing only viral intensity readouts from the experiments might be insufficient. In this project, I describe our computational development as a new alternative approach to improve the reliability for the host factor identification. Our approach is based on characterizing the clustering of infected cells. The idea is that viral infection is spread by cell-cell contacts, or at least advantaged by the vicinity of cells. Therefore, clustering of the HCV infected cells is observed during spreading of the infection. We developed a clustering detection method basing on a distance-based point pattern analysis (K-function) to identify knockdown genes in which the clusters of HCV infected cells were reduced. The approach could significantly separate between positive and negative controls and found good correlations between the clustering score and intensity readouts from the experimental screens. In comparison to another clustering algorithm, the K-function method was superior to Quadrat analysis method. Statistical normalization approaches were exploited to identify protein targets from our clustering-based approach and the experimental screens. Integrating results from our clustering method, intensity readout analysis and secondary screen, we finally identified five promising host factors that are suitable candidate targets for drug therapy.

Secondly, a machine learning based approach was developed to characterize protein-protein interactions (PPIs) in a signaling network. The characterization of each PPI is fundamental to our understanding of the complex signaling system of a human cell. Experiments for PPI identification, such as yeast two-hybrid and

FRET analysis, are resource-intensive, and, therefore, computational approaches for analysing large-scale RNAi knockdown screens have become an important pursuit of inferring the functional similarities from the phenotypic similarities of the down-regulated proteins. However, these methods did not provide a more detailed characterization of the PPIs. In this project, I developed a new computational approach that is based on a machine learning technique which employs the mitotic phenotypes of an RNAi screen. It enables the identification of the nature of a PPI, *i.e.*, if it is of rather activating or inhibiting nature. We established a systematic classification using Support Vector Machines (SVMs) that was based on the phenotypic descriptors and used it to classify the interactions that activate or inhibit signal transduction. The machines yielded promising results with good performance when integrating different sets of published descriptors and our own developed descriptors calculated from fractions of specific phenotypes, linear classification of phenotypes, and phenotypic distance to distinct proteins. A comprehensive model generated from the machines was used for further predictions. We investigated the nature of pairs of interacting proteins and generated a consistency score that enhanced the precisions of the classification results. We predicted the activating/inhibiting nature for 214 PPIs with high confidence in signaling pathways and enabled to identify a new subgroup of chemokine receptors. These findings might facilitate an enhanced understanding of the cellular mechanisms during inflammation and immunologic responses.

In summary, two computational approaches were developed to analyze the image data of the different RNAi screens: 1) a clustering-based approach was used to identify the host factors that are crucial for HCV infection; and 2) a machine learning-based approach with various descriptors was employed to characterize PPI activities. The results from the host factor analysis revealed novel target proteins that are involved in the spread of the HCV. In addition, the results of the characterization of the PPIs lead to a better understanding of the signaling pathways. The two large-scale RNAi data were successfully analyzed by our established approaches to obtain new insights into virus biology and cellular signaling.

Zusammenfassung

Die Untersuchung von Genfunktionen in vielen verschiedenen Behandlungsverfahren, Zelllinien und Organismen wurde durch die Technologie der RNA Interferenz (RNAi) ermöglicht, mit der der Phänotyp von Zellen nach Gen-Silencing bestimmter Gene beobachtet werden kann. In der vorliegenden Arbeit beschreibe ich zwei computergestützte Ansätze, die zur Analyse von Bildern zweier unterschiedlicher RNAi Screens entwickelt wurden. Erstens habe ich einen alternativen Ansatz entwickelt um Host-Faktoren (menschliche Proteine) zu detektieren, die das Viruswachstum sowie die Replikation von Zellen fördern, die mit dem Hepatitis C Virus (HCV) infiziert sind. Verschiedene RNAi Experimente von virusinfizierten Zellen wurden durchgeführt, um diejenigen menschlichen Proteine zu identifizieren, die entscheidend für die virale Infektionseffizienz sind. Trefferlisten aus verschiedenen Laboren haben nur geringe Übereinstimmung gezeigt. Diese Unstimmigkeiten sind möglicherweise nicht nur auf experimentelle Unterschiede zurückzuführen, sondern auch auf die Tatsache, dass die Möglichkeiten der Datenanalyse nicht vollständig ausgeschöpft wurden. Die ausschließliche Betrachtung der experimentell erzeugten viralen Intensitätswerte ist vermutlich unzureichend. In diesem Projekt beschreibe ich unsere computergestützte Entwicklung als einen neuen alternativen Ansatz, um die Verlässlichkeit der Host-Faktor Identifikation zu verbessern. Unser Ansatz basiert auf der Charakterisierung des Clusterings infizierter Zellen. Die Idee ist, dass Virusinfektion durch Zell-Zell Kontakt verbreitet wird oder zumindest durch die Nachbarschaft von Zellen begünstigt wird. Daher betrachten wir das Clustering HCV infizierter Zellen während der Infektionsverbreitung. Wir haben eine Clustering-Detektionsmethode entwickelt, um Knockdown-Gene zu identifizieren, in denen die Cluster von HCV infizierten Zellen reduziert waren. Die Methode verwendet eine distanzbasierte Punktmuster-Analyse (K-function). Der Ansatz konnte signifikant zwischen Positiv- und Negativ-Kontrollen unterscheiden und fand eine gute Korrelation zwischen dem Clustering-Score und den Intensitätswerten der experimentellen Screens. Im Vergleich zu einer anderen Clustering-Methode (Quadrat-Analyse) ist die K-function überlegen. Statistische Normalisierungsmethoden wurden angewendet um Ziel-Proteine aus unserem Cluster-basierten Ansatz und experimentellen RNAi Screens zu identifizieren. Durch Integration von Ergebnissen unserer Analyse, der Analyse von Intensitätswerten und einem sekundaren RNAi Screens, haben

wir schließlich fünf viel versprechende Host-Faktoren identifiziert, die geeignete Kandidaten für eine medikamentöse Behandlung darstellen.

Zweitens wurde ein maschineller Lernansatz entwickelt, um Protein-Protein Interaktionen (PPI) in einem Signalnetzwerk zu charakterisieren. Die Charakterisierung jeder PPI ist elementar für unser Verständnis des komplexen Signalsystems einer menschlichen Zelle. Experimente zur PPI Identifikation, wie z.B. yeast two-hybrid und FRET Analysen, sind Ressourcen-intensiv und daher ist der Rückschluss von phänotypischen Ähnlichkeiten von herunterregulierten Proteinen auf funktionelle Ähnlichkeiten ein wichtiger Aspekt computergestützter Ansätze zur Analyse von umfangreichen RNAi Knockdown Screens. Diese Methoden lieferten jedoch keine detaillierte Charakterisierung der PPIs. In diesem Projekt habe ich einen neuen computergestützten Ansatz entwickelt, der auf einem maschinellen Lernansatz basiert, der die mitotischen Phänotypen eines RNAi Screens verwendet. Der Ansatz ermöglicht die Identifizierung des Wesens einer PPI, d.h. ob sie eher aktivierender oder inhibierender Natur ist. Basierend auf den phänotypischen Deskriptoren haben wir eine systematische Klassifizierung mittels Support Vektor Maschinen (SVMs) etabliert um zu bestimmen, ob ein aktivierendes oder hemmendes Signal propagiert wird. Die SVMs lieferten viel versprechende Ergebnisse mit guter Performanz durch die Integration verschiedener Gruppen von publizierten Deskriptoren und unseren selbst entwickelten Deskriptoren, die aus Fraktionen spezifischer Phänotypen, linearer Klassifikation von Phänotypen und phänotypischen Distanzen zu bestimmten Proteinen berechnet wurden. Ein umfassendes Modell, welches von den SVMs generiert wurde, wurde für weitere Vorhersagen verwendet. Wir haben das Wesen von Paaren von interagierenden Proteinen untersucht und einen Konsistenzwert generiert, der die Präzision der Klassifikationsergebnisse verbesserte. Wir konnten die aktivierende/inhibierende Natur von 214 PPIs in Signaltransduktionswegen mit hoher Sicherheit vorhersagen und identifizierten eine neue Subgruppe von Cheomkinrezeptoren. Diese Ergebnisse tragen möglicherweise zu einem besseren Verständnis zellulärer Mechanismen bei, insbesondere während Entzündungsreaktionen und Immunantworten.

Zusammenfassend wurden zwei computergestützte Ansätze zur Analyse der Bilder der unterschiedlichen RNAi Screens entwickelt: 1) Es wurde ein Clusteringansatz verwendet, um Host-Faktoren zu identifizieren, die entscheidend für eine HCV Infektion sind; und 2) wurde ein maschineller Lernansatz mit verschiede-

nen Deskriptoren angewendet, um PPI Aktivitäten zu charakterisieren. Die Ergebnisse der Host-Faktor Analysen konnten neue Zielproteine aufdecken, die an der Verbreitung von HCV beteiligt sind. Darüber hinaus führen die Ergebnisse zur Charakterisierung der PPI zu einem besseren Verständnis von Signalwegen. Die beiden umfangreichen RNAi Datensätze konnten erfolgreich mit unseren etablierten Ansätzen analysiert werden, um neue Einblicke in die Virusbiologie und zelluläre Signalwege zu erhalten.

Acknowledgements

I would like to take the opportunity to thank many people who contributed to this thesis. First, I would like to thank Prof. Dr. Roland Eils for giving me the opportunity to work in his department and for his support and guidance. Furthermore, I am very grateful to PD Dr. Rainer König for all of his guidance, fruitful discussions, many inspiring ideas, and his outstanding support. I also would like to express my sincere thanks to Prof. Dr. Gerhard Reinelt, Prof. Dr. Reinhard Männer, and Prof. Dr. Hans Georg Bock for their kindness, valuable advice and discussions.

Additionally, I would like to thank all of my colleagues involved in collaborative works through fruitful cooperation, *i.e.*, PD Dr. Karl Rohr, Dr. Nathalie Harder, and Dr. Petr Matula (the BMCV group), Markus Gipp, Dr. Guillermo Marcus (Ziti, Universtät Heidelberg), Dr. Ilka Rebhan and Prof. Dr. Ralf Bartenschlager (the Molecular Virology, Universitätsklinikum Heidelberg), Prof. Dr. Lars Kaderali (Technische Universitt Dresden), Dr. Ellen Ramminger and Prof. Dr. Erich Wanker (Max Delbrueck Center for Molecular Medicine, Berlin). I would also like to thank the Deutscher Akademischer Auslandsdienst (DAAD) for financial support.

Many thanks go to former and current members of the Network Modeling group, Dr. Gunnar Schramm, Dr. Anna-Lena Kranz, Dr. Tobias Bauer, Dr. Kannabiran Nandakumar, Dr. Heiko Mannsperger, Dr. Marcus Oswald, Dr. Rosario Piro, Dr. Zita Soons, Moritz Aschoff, Richa Batra, Ashwin Kumar Sharma, and Volker Ast for their help and warm and friendly working atmosphere. Many thanks go to Rolf Kabbe and Karlheinz Groß for providing an excellent IT infrastructure and Thomas Wolf for fruitful discussion.

I would like to express my special thanks to my parents and my sisters for their love, encouragement, and endless support and my special thanks also go to Kitiporn Plaimas for her continuous encouragement, gentle love and trust in me. They always stand by me in any difficult situation in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The discovery of the RNA interference (RNAi) technology is a major advance in the identification of a specific gene's function or role in signaling pathways. RNAi is a naturally occurring cellular mechanism that permits the silencing of genes and creates phenotypes that can provide clues to the function of these genes. Hence, the technical application of RNAi has been developed on a genome-wide scale and widely used to elucidate central aspects of cell biology. Notably, RNAi technology allows for the analyses of a large variety of different treatments and cell lines, which makes it a desirable approach for large-scale inferences of protein function. Besides this, technologies using fluorescent reporters and imaging by microscopy have been developed for the screening assays and this allows also single cells to be studied over time. Numerous cellular phenotypes (*e.g.*, cell shape, location and signaling response) can be explored from these microscopy assays by using automatic image analysis approaches and numerical features that represent cellular objects are used in pattern recognition, machine learning techniques and statistical analyses for functional analyses [4, 45, 55, 97, 98, 151, 152]. The exploratory data analysis of the RNAi image data has posed challenges and has led to the identification of the function of single human proteins, which correspond to the silenced genes.

RNAi technology can be harnessed to identify drug targets. Several recent studies were reported systematic screening of gene knockdowns to identify host proteins that might support HCV replication [5, 99, 132, 137]. By observing the viral infection after gene knockdown, the silenced genes that reduce the infection rate might be suitable to be targets for drug development. These studies focused on inhibiting host factors (human proteins) instead of viral proteins because human host-factor proteins are evolutionary more stable and will not mutate into drug refractory variants. However, the results from these studies showed only little overlap. This discrepency might result from incomplete data analysis or differences in experimental conditions of these studies. Therefore, an alternative method to improve the reliability of the screens is required. Viral infection is spread effectively by cell-cell contacts. With this mechanism, the clustering of infected cells can be observed during spreading of the infection. To our knowledge an alternative computational method for identifying viral host factors from infected cell localization has not been described earlier. Rather than observing viral intensities which has been used for analyzing infected cell images in traditional way, we developed a computational approach based on a localization analysis of infected cells to identify host factors that might be suitable for therapeutical drug targeting.

In addition, most of the functional processes in a cell involve interactions among proteins. A better understanding of the complex protein-protein interactions can support a better investigation in cell development and disease. To study the interactions of proteins, a variety of high-throughput screens (*e.g.*, the yeast 2-hybrid system [129] or FRET analysis [163]) can be performed to obtain a vast amount of interaction data. However, these approaches can be resource-intensive and infeasible for many protein pairs. Thus, the development of computational approaches for the characterization of protein interactions has become an important pursuit. Besides this, several computational researchers studied protein functions and protein-protein interactions from RNAi screening data using the image processing system, machine learning techniques and statistical analysis in their researches [4, 45, 55, 97, 98]. However, using RNAi technology to better characterize protein-protein interactions has not been performed yet. Identifying if two interacting proteins transduce a rather activating or inhibiting signal can gain a better insight into their cellular function and can be a useful information for pharmaceutical development. We developed an approach based on a machine learning technique to predict the in-

teractions to be activating or inhibiting signals. This approach used features from both published phenotypic descriptors and our own developments calculated from the fraction of phenotypes, linear classification approach from LDA analysis, and distance with protein reference in the network. This integrative approach yielded a more comprehensive model for further prediction.

1.2 Objective and scope

The goal of this thesis is twofold. First, we analyzed the RNAi data of cells infected with the HCV and employed a clustering approach to identify the host factors that are suitable as potential drug targets. The study focused on the clustering behavior of the infected cells after genes were knocked down. The results of our clustering approach were compared with the data from experimental screens to identify potential hits. Second, we analyzed the RNAi data of HeLa cells and developed a machine learning technique for better characterizing known protein-protein interactions. The characterization of protein-protein interactions enhance our understanding of the underlying biological pathways and reveal protein cooperativity that is relevant to disease mechanisms. This study focuses on the similarities between loss-of-function phenotypes of different gene products that are involved in signal transduction pathways.

1.3 Outline of the thesis

Chapter 1 introduces the biological and biotechnical background and further topics related to this thesis and reviews the existing computational methods that concerns to RNAi data analysis, host factor identification and protein-protein characterization analysis. Chapter 2 summarizes the methodologies and datasets applied in this thesis. Detailed descriptions of the methods and algorithms, including the clustering algorithm for detecting a group of infected cells and the machine learning technique for predicting the activities of protein interactions are also provided. Chapter 3 reports the results of identifying host factors involved in viral infection and the results of analyzing the signaling interactions using RNAi screening data. Chapter 4 provides the summary, discussion and outlook of this thesis.

1.4 Publications

- **A. Suratanee**, I. Rebhan, P. Matula, A. Kumar, L. Kaderali, K. Rohr, R. Bartenschlager, R. Eils, and R. König, Detecting host factors involved in virus infection by observing the clustering of infected cells in siRNA screening images. *Bioinformatics*(Oxford) (26) 18, 2010.

- M. Gipp, G. Marcus, N. Harder, **A. Suratanee**, K. Rohr, R. König and R. Männer. Haralick's texture features computed by GPUs for biological applications. *IAENG International Journal of Computer Science*, 36(1), 2009.

- M. Gipp, G. Marcus, N. Harder, **A. Suratanee**, K. Rohr, R. König and R. Männer. Haralicks texture features computation accelerated by GPUs for biological applications, *4th International conference on high performance scientific computing, Modeling, Simulation and Optimization of Complex Processes*, Hanoi, Vietnam, 2-6 March, 2009.

- M. Gipp, G. Marcus, N. Harder, **A. Suratanee**, K. Rohr, R. König and R. Männer. Accelerating the computation of Haralick's texture features using Graphic Processing Units (GPUs). *Proceedings of the world congress on enginerring*, London, U.K., 2-4 July, 2008. Newswood Limited, International Association of Engineers.

The results of our research about the host factor identification have been published in the journal *Bioinformatics* [133]. The manuscript covering the characterizing signal interaction part is currently in preparation. A part of this project is involved in a publication [47] published in *IAENG International Journal of Computer Science*.

1.5 Biological and technical background

In this section, I briefly summarize the biological and biotechnical background concerning the application of my work. First, I briefly describe the process of RNAi to promote the understanding of the data I analyzed. Next, I give a short overview of the high-throughput RNAi screening technique used to generate the RNAi data. Thereafter, I provide a short overview describing the biology of HCV. Finally, I give an overview of signal transduction and protein-protein interactions.

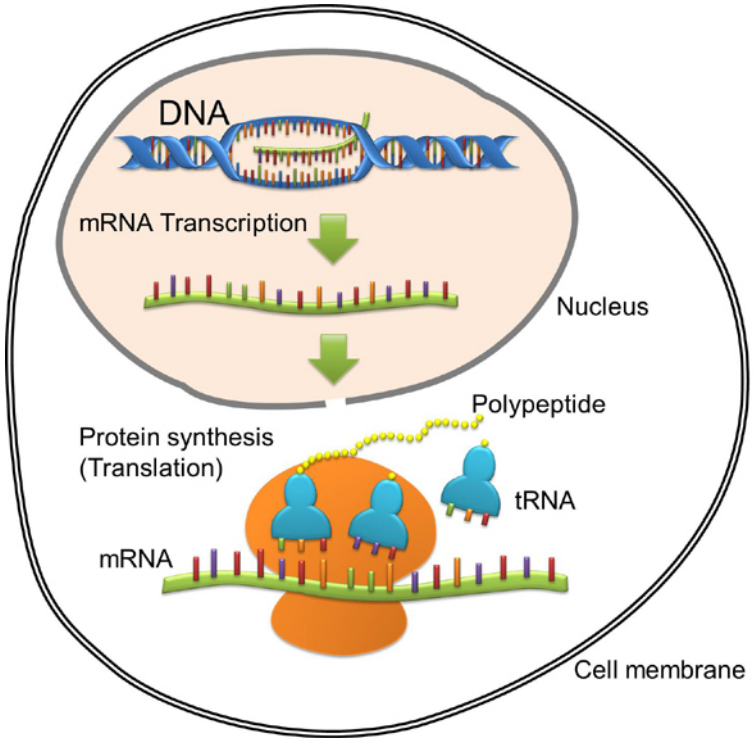

Figure 1.1: DNA to protein. A complementary RNA copy is created from DNA sequence during transcription. The genetic information is carried by the mRNA, which is used to synthesize the protein molecules on the ribosomes. (The figure is modified from http://de.wikipedia.org/wiki/MRNA).

1.5.1 RNA interference

A common approach used to discover the function of a gene is to down-regulate the expression of the gene, which down-regulates the corresponding protein. The phenotypic effects caused by this down-regulation are then studied. The discovery of RNA interference (RNAi) or gene silencing using double stranded RNA has allowed the disruption of expression. This RNAi is a cellular mechanism of post-transcriptional gene silencing to prevent the cell from expressing foreign genetic material, *e.g.*, genetic material from a virus. In the nucleus of a normal cell (Figure 1.1), the DNA sequence of a gene is used as a template to synthesize the ribonucleic acid (RNA) molecules during the processes of transcription. A protein-coding gene that is copied into an RNA molecule is further processed into a messenger RNA (mRNA). The mRNA is then transported into the cytoplasm and binds to

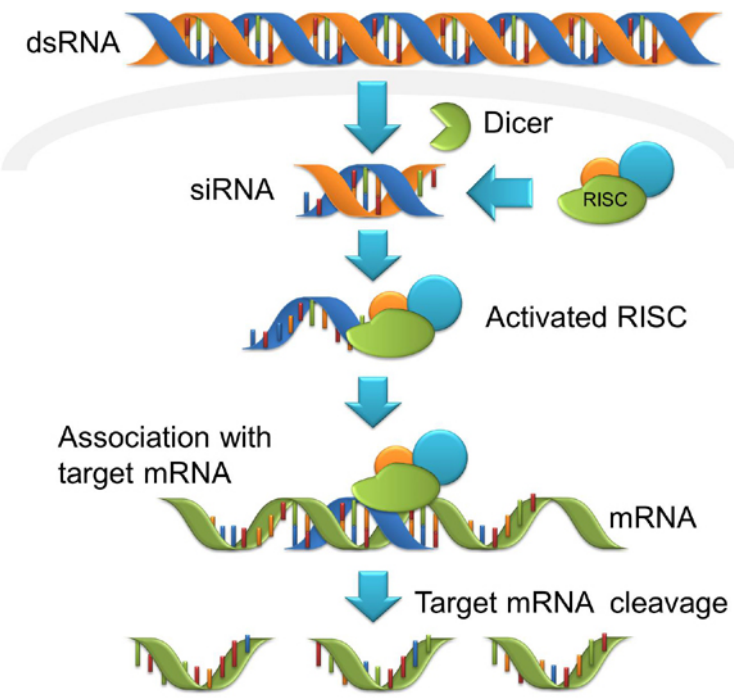

Figure 1.2: RNA inference process. An enzyme of the Dicer family cleaves the RNA into small pieces, called the siRNAs. The siRNA activates the RISC and aids in the recognition of the complementary mRNA. The mRNA is cleaved and destroyed; thus, the corresponding protein cannot be produced. (The figure is modified from http://www.scbio.de/gene_silencers.html).

a ribosome that translates the mRNA and produces the respective protein. The mechanism of RNAi (Figure 1.2) initiates from a double stranded RNA (dsRNA) in which one strand is complementary to a section of the mRNA. An enzyme of the Dicer family proceeds to cleave and cut the RNA into small pieces called the small interfering RNAs (siRNAs). Then, one strand of the siRNA called the antisense strand becomes the 'guide', and the other strand becomes a temporary 'passenger', which is quickly degraded [37]. The antisense strand (guide) is integrated into an RNA-induced silencing complex (RISC) and then forms the activated RISC. The antisense strand aids the RISC complex in the recognition of the complementary mRNA, which it cleaves and systematically destroys the cognate RNA. The respective protein cannot be produced after destruction of the mRNA by this RNAi

process. To silence a gene, the siRNAs that have base-sequence complementarity to the mRNA of the target gene are transfected into the cell.

1.5.2 High-throughput screening of the RNAi experiments

When laboratory automation (*e.g.*, tissue culture facilities, arrayer robots, and plate reader), software control, and computing infrastructure are employed, the use of high-throughput screens allows millions of genetic tests to be rapidly conducted. The application of RNAi technology to a high-throughput screen is a powerful method to address many questions of cell biology. This loss-of-function screen is also particularly useful for the analysis of signal transduction pathways [37]. In high-throughput screens that use genome-wide siRNA knockdown experiments, approximately 22,000 human genes can be screened. To attain this number, the experiments can only be performed after the optimization and automation of the experimental processes. In 384-well plates, the siRNA-gelatine transfection solution is prepared and then arrayed into single-wells of the LabTek cover glass live cell imaging dishes. The spot diameter is approximately 400 m, and the spot-to-spot distance is approximately 1125 m. These siRNA microarrays are dried and stored overnight. After drying, the HeLa-H2B-GFP cells are plated on the microarrays and transfected by growing these cells on the siRNA spots. Images are acquired with an automated microscope every 30 min for 44 hours. The imaging starts 20 hours after plating the cells on the siRNA microarrays. To image as many microarray spots as possible within a time lapse of 30 min, the number of spots that can be imaged simultaneously, the time spent at each spot and the desired temporal resolution have to be carefully manipulated. Additional details of genome-wide high-throughput screens are also available (*e.g.*, [97, 98]).

1.5.3 Hepatitis C Virus

Approximately 170 million people are infected with the HCV worldwide [109]. The HCV is a major cause of persistent chronic infections that lead to development of steatosis, liver failure, liver cirrhosis and hepatocellular carcinoma [44]. The HCV is a single-strand RNA virus that has an average incubation period of 6-8 weeks. The HCV infection is often asymptomatic, and hence the detection of the HCV at an early stage is difficult. Therefore, the HCV is often referred to as a “silent

disease" [123]. The mechanism of the HCV life cycle is still unclear. A decade ago, the model of viral entry was developed. It was found that a key cellular protein for viral entry, CD81, is needed and binds to the viral structure protein E2 on the surface of the HCV caspid [108]. Many more proteins were subsequently identified as factors involved in the HCV entry, including two essential proteins, SR-BI and claudin-1, as well as, accessory factors, such as glucosaminoglycans and low-density-lipoprotein receptors (LDL receptors) [34, 42, 107, 121]. Generally, the viral envelope protein of the HCV plays a central role in the HCV binding to host receptors and membrane fusion. After the HCV uncoating, the viral genome is then translated in preparation for viral replication. The translation of the viral genome generates nonstructural and structural viral proteins, which are needed for the viral replication and assembly of new viral particles. Viral replication is carried out in a convoluted membrane structure called a membranous web. The newly synthesized viral RNA strand is released from the membranous web and passed to the core protein via the NS5A. The core protein is translocated onto the surface of a lipid droplet or an endoplasmic reticulum (ER) membrane for efficient formation of the viral particles, and then encloses the synthesized viral genome to form a capsid. The capsids are enclosed by an endoplasmic membrane containing the viral envelope proteins and are then released into the ER lumen. Finally, the viral particle is released from the infected cells [95, 112]. The HCV has several mechanisms it employs to inhibit the host response. The HCV infection induces an interferon response in the liver of patients, and the expression of several HCV proteins has been shown to inhibit and evade the innate antiviral response of host cells [65]. Recently, Moriishi and coworkers reported that a HCV core protein cooperates with the host factors and causes the lipid alternation, oxidative stress, and the progression of cell growth. To maintain efficient viral replication and production, other viral proteins interact with the host proteins, including molecular chaperones, membrane-anchoring proteins, and enzymes associated with lipid metabolism [95]. Hence, the investigation of host factors is progressing, and this progress is crucial for the discovery of treatments for the HCV.

1.5.4 Signal transduction and protein-protein interaction networks

Cells typically receive chemical or mechanical cues from their environments. In response to these cues, the cells send and propagate signals through signaling cascades. There are many types of these pathways and they are commonly categorized as metabolic pathways, gene regulatory pathways, or signal transduction pathways. The metabolic pathways are well-studied and comprise a series of biochemical reactions that maintain all cellular processes. To produce cellular energy or synthesize cellular components, the metabolic pathways break down large nutrient molecules (*e.g.*, proteins, carbohydrates and fats) into small molecules. The gene regulatory pathways or transcriptional regulatory pathways concern transcription factors, their respective target genes, and the regulation between. Transcription factors bind to the DNA at specific binding regions to stimulate or repress gene transcription, and this binding regulates the production of the corresponding proteins. The signal transduction pathways connect extracellular signals and transcription factors by a complex system of interactions between signaling molecules within the cell. In a typical signal transduction pathway, a receptor is a protein on the cellular surface that receives and responds to a stimulus. An intracellular response is initiated after the signal interacts with the receptors. The resulting message is transmitted by specific proteins that trigger a specific action in the cell. Most of our understanding of cellular processes is based on the identification and characterization of the interactions between proteins and other biomolecules. These protein interactions propagate the signal, which is the main process of signaling transduction. In this thesis, I will use the term protein-protein interaction (PPI) to refer to a physical interaction between proteins. Examples of PPIs include the phosphorylation, binding, and association of proteins to forming protein complexes. The second part of this thesis characterizes the PPIs of signaling pathways.

1.6 Existing computational approaches

In this section, I review various existing computational approaches of analyzing RNAi screening data that concerns to our works, *i.e.*, the hit identification analysis and protein-protein characterization analysis. I firstly explain existing approaches

for analyzing the RNAi microscopy image data. Next, I describe existing approaches for identifying host factors being important for viral infection. Finally, I explain the existing approaches for inferring protein functions and protein-protein interactions on a large scale.

1.6.1 Computational analysis of the RNAi image data

The use of fluorescence microscopy for the imaging of RNA interference (RNAi) knockdown screens has become a preferred method to identify the protein function of silenced genes and can be harnessed to detect potential drug targets. Computational approaches for automatic analysis of cell microscopy images after knocking down genes have been successfully developed to describe the loss-of-function morphological features. A goal of the RNAi knockdown screen is to study the effects of experimental treatments on a cell population by comparing population-based features. Usually, cell nuclei are the main labeled compartments of interest. Also other subcellular structures, *e.g.*, the cytoplasm, cytoskeleton, or proteins indicating a specific cellular response (such as from virus infection) have been additionally labeled in separate channels [82, 92, 135, 153]. Studies were reported that investigated single cells using phenotypic features such as cellular area, diameter, eccentricity, texture, granularity, moment [18, 56, 64, 75, 98, 150]. The main steps for analyzing the data consisted of (1) segmentation, (2) feature extraction, and (3) classification. The aim of segmentation is to identify the cells in the images. The image is seperated into different regions, each containing a single cell and the cells are seperated from the background. This procedure can be done by several segmentation algorithms, *e.g.*, threshold- or edge-detection-based algorithms. For example, the goal of the Otsu thresholding technique [100] is to find the optimal threshold that separates the pixels into two populations (the cells and background) by minimizing the in-class variance and maximizing the between-class variance. After each single cell segmentation, cellular features are computed based on the identified single cell. The features of a single cell are a numerical vectors representing the sizes, shapes, or textures of a cellular object. These features can be analysed directly by comparing the distributions of the features between two different experiments, *e.g.*, normal versus cancer nuclei, using statistical tests [87].

Additionally, when analyzing the cellular phenotypes, we wanted to observe only specific cellular shapes occuring, *e.g.*, during apoptosis or mitosis. In this case,

machine learning techniques are required for learning the specific phenotypes and performing the phenotypic prediction. Based on the extracted features and labeled classes, machine learning techniques have been applied to classify cell nuclei into different cell cycle phases [46, 97, 98, 151]. Held *et al.* [58] used an SVM technique to classify cells into interphase, six mitotic stages and apoptosis based on 186 quantitative features describing texture and shape. Neumann *et al.* [97] used a live-cell assay to profile the cell-division of the HeLa cells. They silenced each of 21,000 protein-coding human genes in a separate cell population using the RNAi method and observed the effect using fluorescently labeled chromosomes that express histones (H2B) tagged with the green fluorescent protein (GFP), followed by automated high-throughput time-lapse microscopy. They also used SVMs for classifying the cell nuclei into several classes, *e.g.*, interphase, mitosis, apoptosis, clustered nuclei, and artefacts, based on 214 extracted features of texture and shape. Harder *et al.* [54, 57] extracted 376 features based on the size and shape, texture (Haralick features), geometric moments, and granularity to classify imaged life cells into 12 classes of cell division cycle phases. A variety of applications followed extracting the texture of RNAi transfected cells from large-scale cellular phenotypic assays, and using machine learning methods allowing the classification of cells to identify subcellular location [27, 104] and specific cellular features (*e.g.*, the mitotic state and viability of the cell) [18, 55, 56, 64, 75, 98, 150]. The classified phenotypes were used for further analyses, *e.g.*, to put-up models for cell division cycles or progression of mitotic events [57, 97, 98]. Apart from the cell nucleus staining, also for additional other cellular components these method were applied and the morphological features were extracted using the fluorescent signal intensity [52, 92, 135] of the structure of interest, *e.g.*, of the mitotic spindle, centrosomes, or spliceosomes [48, 105]. Matula *et al.* [92] analyzed the RNAi of cells infected with GFP expressing HCV. They measured the viral intensity from the GFP channel in cytoplasm and used the intensity to compute the infection rate and classify each single cell into two classes (infected and non-infected cells). The classification was performed by finding the optimal intensity threshold that maximize the difference of the infection rates between the positive and negative controls on a labtek. The target gene was identified from the infection rate or from the average of viral intensity.

From the aboved studies, applications of the classification methods of morphological phenotypes were mostly used for finding sets of functionally related genes

that showed similar knockdown phenotypes. We exploited the classification approaches for identifying infected cells to our new approach to identify human host factors. Moreover, the methods for cellular phenotype classifications were exploited for developing phenotypic descriptors related to calculation of phenotypic fraction or phenotypic similarity to characterize activities of a protein-protein interaction.

1.6.2 Computational approaches for targeting host factors

Despite many substantial discoveries in virology, viruses remain a major cause of severe diseases including Dengue fever, hepatitis, immune deficiency and severe influenza. Viruses employ specific human host proteins (*i.e.*, host factors) for each step of their 'life' cycle [19, 88, 91]. Discovering these host factors may not only unravel the fundamental principles underlying the mechanisms of viral action (*e.g.*, viral replication), but also, notably, may lead to promising drug therapies that are not affected by the high mutational variability in viral populations. Computational approaches to determine human proteins involved with virus or other pathogens mostly exploit information from available pathogen-host, protein-protein interaction, or gene ontology databases [1, 69, 164]. The similarities or interactions of host and pathogen sequences, structures, or domain-interactions are employed for finally predicting the probability of the interactions between host proteins and the pathogen [31, 33, 36]. Doolittle *et al.* [33] developed a computational approach for predicting host factors for Dengue virus (DENV) for both the host (human) and the vector (insect). The approach was based on the similarity of 3D protein structures of DENV proteins and human proteins (hDENV-similar protein). They investigated the interactions of hDENV and other target human proteins in the protein-protein interaction database (the Human Protein Reference Database, HPRD) and predicted that the target proteins might also interact with the DENV protein. However, these methods based structure similarity and lack of predictions for viral proteins which do not have a human homologous structure. Moreover, pathogens including viruses show high variability and can evolve exploit the host proteins using various strategies as well as effective escape machanisms [71]. Therefore, the network of virus and human proteins are clearly dynamic and undergo diverse mechanisms of actions [71, 134]. Hence, these methods need to be further improved. Rather than using the pathogen-host interaction analysis, RNAi knockdown screens have been used and we propose a new technique for detecting host factors in this RNAi data

in this thesis.

There are several genome-wide and more selected screens like *e.g.*, kinase siRNA screens using an infectious HCV to be brought into a cell culture to identify host factors [77, 137, 142, 145]. One goal of these RNAi screens is to directly identify the siRNAs (hits) that generate meaningful cellular phenotypes of cells infected with the virus. Recently, Reiss and colleagues [115] performed an siRNA screen of the human kinome to detect the host factor requirements for HCV replication. They identified 13 different kinases that are required for HCV replication. However, most of the previous results show the difficulty to obtain a consensus set of gene targets. Li *et al.* [77] and Randall *et al.* [114] used siRNA against human host factors using the same HCV genotype and modeled the human cell system. The results showed only eight genes that overlapped across all platforms. This discrepancy might result from incomplete data analysis or differences in the experimental conditions of these studies, such as the use of different viral strains, time intervals or silencing sequences. Brown *et al.* [14] showed the results of a computational analysis of genes identified from four different experiments (protenomics, mRNA microarray, RNA-Seq, and siRNA). They studied the effects of infecting cells with HCV by measuring changes of infection. They performed pathway enrichment tests using GeneGO™ Metacore™ and revealed a greater overlap at the pathway level. They found 16 pathways which were significantly enriched in three out of four experiments. These pathways are known to be modulated by HCV infection. Therefore, this finding showed that the development of an alternative approach might support to get insight from the experimental data.

However, most of the studies for host identification are performed by the RNAi experiments in the wet-lab and the results are analyzed with statistical approaches [9, 77, 114, 115, 137, 142, 145]. The statistical analysis for hit identification is a bioinformatic approach that is an important step after conducting the knockdown experiments to recover the set of important genes. To allow a comparison of the data from different plates (labteks) and positions on these plates, data normalizations need to be conducted. A variety of current normalization approaches have been developed to analyze the RNAi screens [89]. For the microscopy-based screens, the mean or median fluorescence intensities of the cells in a spot are calculated. These summarized values are used to normalize within and between different experiments [7, 9, 115, 154]. For the RNAi screen of cells infected with a virus (*e.g.*, the

HCV or Dengue Virus), the viral expression of the GFP (intensity) is analyzed based on the average fluorescence intensity per spot [115]. The normalization can be performed using controls, such as siRNA controls with random, non-functional sequences; this normalization is called a control-based normalization. In contrast, the normalization can also be performed using all measured data as a control [7]. The normalization of the data from screens can be performed using the percentage of control method, the normalized percent inhibition method, the z-score normalization or the B-score normalization [7, 89]. The z-score is a measure of the standard deviation away from the mean. The z-score is frequently used to normalize data of high-throughput cell array screens; however, this method is sensitive to outliers. It was suggested that the B-score normalization is a robust application of the z-score normalization [89]. The B-score normalization accounts for row and column variations and has the advantage that it minimizes the biases from positional effects. To reduce the row and column effects, B-score normalization uses a two-way median polish procedure, which is an iterative algorithm that alternates row and column operations. By using the medians rather than the means, the B-score normalization is less affected by the presence of outliers. On each iteration of the two-way median polish procedure, the row median, column median, and median of these medians are computed and accumulated systematically into the row effects, column effects, and an overall level effect. This procedure is continued until the value of the row and column medians nears zero. The two-way median polish procedure is performed for each plate. To account for plate-to-plate variablility, the resulting residuals of each plate are then divided by their median absolute deviation (the median of the absolute deviations of the medians) from all the residuals of the plate.

After the normalization, the data are processed to determine which genes differ significantly from those of the negative controls, which identifies the hits or positives from the screen. Screeners might simply select a discrete number of top scoring genes from the screen as the hits. However, many hit identification techniques are available to obtain the quality hits and reduce the risk of false positives. The hit identification can be performed with the mean $\pm$ k standard deviation or median $\pm$ k median absolute deviation. The genes are identified as hits if they surpass these thresholds. The z-score normalization is simple and frequently used for the hit identification. However, a robust z-score is preferred for hit identification. The robust z-score is computed by subtracting the median instead of the mean from the measured values

and then dividing by the median absolute deviation. The genes with low, significant z-score values (P-value$<$0.05) are selected as the hits [9, 117].

From the above reviewed studies, the host factors were identified from the statistical analyzing of the intensity readout of GFP-expressing viruses from knockdown experiments or inferred from the pathogen-host analysis. However, there are problems regarding inconsistency of target lists and complexity of dynamic host-virus networks. Therefore, an alternative approach is required to identify human target genes that improve the reliability for host identification. In this thesis, we developed a new approach based on cell localization to identify the host factors on the screens and used the above statistical techniques to support the hit identification.

1.6.3 Inferring protein functions and protein-protein interactions

The accurate reconstruction of signal transduction pathways within cells is central to elucidating the cellular mechanisms of pathogenesis. The interactions within signaling cascades are often specific to a given treatment or disease under investigation [74]. With the help of manual curation, the experimental validations of direct PPIs and functional relationships have been extracted from the literature and assembled in well-established databases [66, 68, 74, 128]. To identify new PPIs, a vast amount of interaction data has been assembled from a variety of high-throughput screens, including data from the yeast 2-hybrid system [129]. These screens can be resource-intensive, especially if any possible interaction needs to be experimentally investigated (*e.g.*, 12.5 million experimental interaction assays for a selection of 5,000 proteins). Therefore, the use of computational approaches is suggested for the statistical inference of the PPIs using information from the co-expression of genes, co-evolutionary studies and natural language processing (STRING [63, 136]). Bakal *et al.* [4] used neural networks that based on morphological features of cells to infer functional similarity from phenotype similarity of a smaller set of genes with well-characterized functions. They first calculated 145 morphological features for each cell to identify 7 classes of morphology that based on known phenotypes and this was the result related to the perturbation of key signaling molecules (*e.g.*, Rho and Rac). They trained a set of artificial neural networks to identify these phenotypes. The result was a matrix with seven columns; each represented the similarity

score of the cell to the phenotypic category. They further performed a hierarchical cluster analysis that allowed identification of the local signaling network with functional characteristics regulating cell shape and migration. Fuchs *et al.* [45] described an experimental and computational approach to predict gene functions basing on changes in the morphology of individual cells within cell populations. They assessed the effect of siRNA transfection on HeLa cells which DNA and the cytoskeletal proteins actin and tubulin were stained. 51 morphological features were computed and applied to SVMs. The classification results were employed to generate 13 phenoprints used to compute similarity distances. The clustering of genes was performed based on the similarity distances and elucidated new functions of genes that were involved in the organization of the spindle. Neumann *et al.* [98] analyzed time-lapse microscopy siRNA data to identify a set of genes involved in cell division, migration and survival. They computed about 200 morphological features for each single cell and classified them into 16 phenotypic classes using SVMs. A phenotypic profile of each class was computed that based on the time-lapse image sequence. They performed hierarchical clustering of genes by their phenoprints in all morphological classes, taking both the temporal change and the severity of the phenotype into account to identify a group of mitotic genes. The rationale of these approaches is that similar cellular phenotypes will arise if the functions of the knocked-down genes are tightly linked. For example, cellular phenotypes are expected if the proteins corresponding to the knocked-down genes are part of the same protein complexes or are mutually dependent for the propagation of signals. Recently, Vinayagam *et al.* [149] developed an experimental and computational approach predicting the directionality of signal flow in a signaling network. They initially generated a PPI network from yeast two-hybrid data and combined the information with publicly available interaction data that resulted in a network comprising 1126 proteins and 2626 PPIs. The method predicted the flow of signaling cascades from membrane receptors to transcription factors with the shortest path connections. They used a naive Bayesian classifier for the prediction with 8 probability features of the direction between the proteins and yielded a good performance. Although the approach was able to predict the signal flow, the study of the sign (activation and inhibition) of the signal transduction has not been addressed.

Investigating how signal transduction by PPIs is mediated by phosphorylation is an alternative way to gain insight into intracellular signal transduction. The

goal of phosphorylation studies is to understand the nature of these phosphorylation interactions by mapping the phosphorylation sites and effector kinases. This mapping aims to reconstruct a cellular signaling network [140]. The knowledge of phosphorylation site prediction could be used for interpreting activating protein pairs that a protein in the pair is a protein kinase that could be an evidence for a protein-protein activation. Protein phosphorylation is a post-translational modification of proteins that affects approximately one-third of all cellular proteins [26]. Both experimental and computational approaches have been developed for phosphorylation site detection. The software or databases for this detection have been provided [125, 139, 140, 158, 159, 169]. Tan *et al.* [139] developed a sequence alignment approach to reconstruct conserved kinase-substrate networks. They identified proteins that were tightly regulated by phosphorylation. Using topology features of a predicted human phosphorylation network, a regulatory hub protein was found to be highly phosphorylated, and the identified proteins were evidenced to be associated with various diseases, *e.g.*, diabetes, cancer, or Alzheimer's disease [139].

From the above reviewed studies, the aim of the study from Vinayagam *et al.* [149] is closely related to our work. Most of the above studies addressed approaches for inferring protein function or protein-protein interaction. However, they do not address the question about the activities between the interacting proteins in signaling transduction. The knowledge from predicting kinase-specific phosphorylation site as mentioned above can support the activation interactions from our study.

1.7 Main contributions of this thesis

In the following, I summarize the main contributions of this thesis:

- **Host factor identification in cells infected with the HCV**
  In this study, we investigated the HCV infection in a human hepatoma cell line to detect human host factors that are necessary for viral infection. A comprehensive set of 719 genes expressing different kinases was screened by employing the RNAi technology [115]. We developed a computational approach basing on a well-known point pattern analysis approach, the K-function, to detect clustering of the cells (see Section 2.1.4). This approach observed a reduction of viral infection in a reduced grouping (clustering) of the infected cells. For each

knockdown experiment, we compared the clustering of the infected and non-infected cells and estimated the reduction in clustering of the infected cells. We also applied an alternative clustering method, the quadrat analysis, to measure the infection phenomena of the cell distribution (see this method in Section 2.1.4 and the results in Section 3.1.1). Different bioinformatics approaches were applied to the data to identify the host factors that significantly reduce the HCV infection efficiency. We employed a statistical method described recently using B-score and z-score normalization of the intensity readouts from the segmented cellular images [13], the intensity readouts of a luciferase based secondary screen and our clustering scores (see Section 2.1.6). We yielded 30 promising candidates suiting as potential host factors for therapeutical drug targeting. Five of these candidates were found using all three methods: the CD81, PI4KA, CSNK2A1, SLAMF6 and FLT4 (see Section 3.1.1). In conclusion, we report an alternative method for high-throughput imaging methods to detect host factors being relevant for the infection efficiency of viruses. This method is generic and has the potential to be used for a large variety of different viruses and treatments being screened by imaging techniques.

- **Characterization of signaling interactions**
  The aim of our study was to elucidate if two interacting proteins positively propagate a signal (activation) or if their interaction rather lead to a conversion of the original signal (inhibition). For this, we developed a workflow that employed a machine learning approach based on the idea that activating signals lead to similar knockdown phenotypes of the respective interacting proteins, whereas the inhibitory signals lead to rather dissimilar phenotypes (see Section 2.2.1). We used a large range of phenotypic descriptors calculated from fractions of specific phenotypes, linear classification of phenotypes, and phenotypic distance to distinct proteins (see Sections 2.2.5 and Section 3.2.2). We applied this approach to cellular images collected in the Mitocheck genome-wide RNAi knockdown screen [98]. Support Vector Machines were employed for the interaction classification (see Section 2.2.6). With this, characterizations of protein-protein interactions were identified. The results from classifications showed that our methods can be used to classify interactions as having a role in the activation or inhibition of signal transduction with AUC of 0.76 (see Section 3.2.3). Consistency score was established for investigat-

ing the nature of pairs of interacting proteins (see Section 2.2.9). This score used the performance criteria of the machine learning method to estimate the consistence of pairs of individually knocked-down genes. The relevance of our consistency scores was validated using other independent databases through GSEA enrichment tests (see this method in Section 2.2.10 and the results in Section 3.2.4). In a case study, we analyzed the signal transduction pathways leading from the cytokine receptors to the transcription factors that were known to be controlled by these pathways.

Chapter 2

Methods

This chapter is divided into two parts. Section 2.1 describes the methods used to detect the host factors necessary for viral replication and the analysis of the clustering of cells infected with HCV. The clustering approach, which is based on the spatial distance, and the statistical analysis for defining hits are described. Section 2.2 describes the general workflow and methods used for the characterization of the PPIs as having a role in the activation or inhibition of signal transduction and includes a detailed description for generating the phenotypic descriptors. The classification approach and the measurement of its classification performance are also explained.

2.1 Clustering of cells infected with Hepatitis C Virus

2.1.1 General concept and workflow

The detection of human proteins that are involved in viral entry and replication is facilitated by the modern high-throughput RNAi screening technology. However, the hit lists from different laboratories have shown only little consistency. This lack of agreement might result from experimental discrepancies or unexplored possibilities in the data analysis. We would like to improve the reliability of the RNAi screens by combining a population analysis of infected cells with an established dye intensity

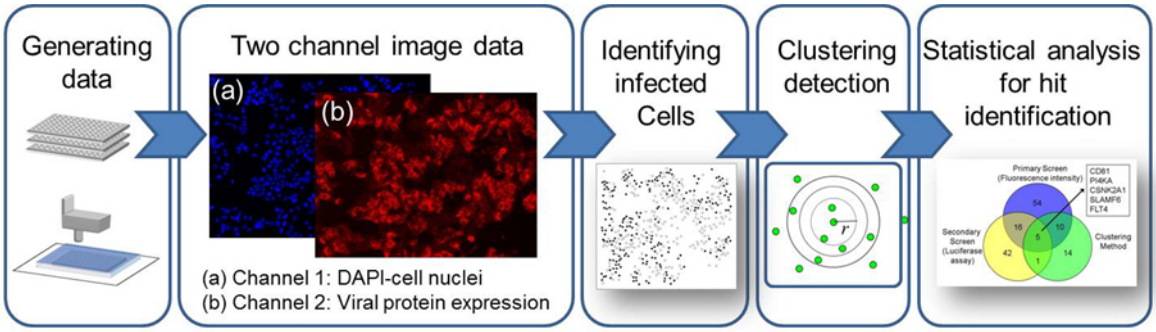

Figure 2.1: The workflow used to identify the host factors that are crucial for viral infection efficiency. The RNAi data of cells infected with HCV are generated as two channel image data (the DAPI-stained cell nuclei and the GFP expressed by the virus). An automated imaging system is employed to identify the infected and non-infected cells. A clustering approach is then performed to analyze the clustering of infected cells. Statistical methods are applied to identify potential hits by comparing the hits from the clustering approach with the hits from a standard procedure and a secondary screen.

readout. The viral infection is mainly spread by cell-cell contacts, and the clustering of infected cells can be observed during spreading of the infection *in situ* and *in vivo*. We employed this clustering feature to define the knockdowns that harm the viral infection efficiency of the human HCV. Images of the cells that were knocked down for 719 human kinase genes were analyzed with an established point pattern analysis method, Ripley's K-function. This method was used to detect knockdown cells in which the viral infection did not show any clustering and therefore were hindered to spread their infection to the neighboring cells. The results were compared in a statistical analysis that used intensity readouts of the GFP-expressing viruses and a luciferase-based secondary screen. Five promising host factors were identified and are suitable as potential targets for drug therapy.

An overview of our workflow is shown in Figure 2.1. During screening, images were taken under fluorescence microscopy of the infected human cells with knocked-down genes. The cells were cultured and treated on printed plates, and the siRNA and transfection reagents were spotted on a chamber plate at known locations in a grid pattern. Only the cells located within the area of a printed spot took up the corresponding siRNA and underwent gene silencing. The two-channel images were

acquired using an automated fluorescence microscope. The first channel displayed DAPI stained cell nuclei. The second channel represented a viral expressed fluorescence protein (GFP, Green Fluorescence Protein). An automated system, which was described in detail by Matula and coworkers [92], was employed. In the DAPI channel, the single-cell nuclei were segmented, and the viral protein production levels (viral signal) of each cell were computed by the mean intensity in channel 2. According to the viral signal, the cells were classified as infected or non-infected based on a thresholding procedure. The cells with a viral signal that was less than the threshold were classified as non-infected, otherwise cells were classified as infected. The threshold was defined by maximizing the difference in the infection rates between the positive and negative controls, which were spotted on the same plate. We applied the K-function to the spot distributions using the local spatial variation, which is a statistical clustering method. We found candidate host factors that are suitable for therapeutical drug targeting.

2.1.2 Data source

The experimental data were generated by our collaborator, Ilka Rebhan at the Department of Molecular Virology, UniversitätsKlinikum Heidelberg. The siRNA library used for the primary screen of this study was purchased from Ambion (*Silencer*® Human Kinase siRNA Library V3 (AM80010V3)). The reverse transfection of the siRNAs into Huh7.5 cells [8] in a LabTek format was optimized according to a previously described protocol [39]. Overall, 2157 siRNAs targeting 719 human kinase genes plus positive controls targeting the entry receptor CD81 or the viral genome itself (HCV321 and HCV138) and four different negative controls (non-silencing siRNA) were spotted in transfection mixture onto LabTeks. After the seeding of the Huh7.5 cells, we allowed the siRNA silencing to occur for 36h. The cells were infected with a HCV GFP reporter virus, fixed 36h later and immunostained with a GFP-specific antibody. The cellular arrays were imaged with a scanning microscope (Scanˆ R, Olympus Biosystems) using the 10x objective (Olympus, cat. no. UPSLAPO 10x), and images were analyzed with an image analysis method (see Section 2.1.3). The primary screen was repeated in 12 times. All images with less than 125 or more than 500 cells within the siRNA spots were excluded from the analysis. As an additional quality control for staining artifacts, all images were analyzed by eye; this quality control step resulted in the exclusion of 15% of the images. Statistical

analysis was performed to compute a mean z-score and a P-value for each gene to facilitate the selection of candidate genes (Section 2.1.6). During the validation of 178 gene candidates selected from the primary screen, three independent siRNAs per gene were used to minimize the number of potential off-target hits. In addition, the format of the assay was changed to a statistically more robust 96-well plate format to increase the number of transfected cells per siRNA and thus the statistical power of the assay (approximately 300 cells in the LabTek format but approximately 10,000 in this well-based assay). The solid phase method of reverse siRNA transfection was adapted to the 96-well plate format as described by Erfle and coworkers [40]. Briefly, the siRNAs are printed together with a gelatin solution at defined locations on the glass slides. After drying the wells, the substrates can be stored for up to 15 months without any loss in efficacy or directly used for knockdown studies. This method is called a "reverse transfection" because the order of addition of the siRNAs or expression plasmids and cells is reversed in comparison with the conventional transfection method [39, 40]. This assay format allowed the use of a luciferase reporter virus that also facilitated the analysis of the screen. To validate the effects of the kinase knockdown experiments on the HCV entry and replication, 5 x 10^3 Huh7.5FLuc cells (stably expressing firefly luciferase) were seeded in each siRNA-coated well of a 96-well plate. After 36h, the cells were infected with a HCV renilla luciferase reporter virus. Forty-eight hours post-infection, the cells were harvested, and the firefly luciferase and renilla luciferase activities were measured. The secondary screen was performed twice in duplicates.

2.1.3 Image analysis of HCV infected cells

To analyze the images of the siRNA screen, an automated system, which is described in detail elsewhere [92], was employed. Briefly, the inputs of this system consisted of two dye-channel images from a chamber plate with printed siRNA spots. The fluorescence signals originated from the DAPI-stained cell nuclei (1^{st} channel) and Green Fluorescence Protein (GFP), that incorporated into the viral strain (2^{nd} channel). In the DAPI channel, the single-cell nuclei were segmented using an edge-based approach that combined the responses of the gradient magnitude and the Laplacian of the Gaussian filters with the morphological closing and hole filling operators. The nuclei were identified among the segmented objects by applying the size, intensity, and circularity criteria. The viral protein production level (virus signal) of each cell

was computed using the mean intensity in channel 2 inside the nucleus neighborhood. The positive and negative controls were spotted on each plate. In the positive controls, the siRNAs hindered viral protein production, which resulted in a low viral signal; whereas, in the negative controls, the viral replication was unaltered. According to the viral signal, the cells were classified as infected or non-infected using a thresholding concept. The cells with a viral signal less than the threshold were classified as non-infected, otherwise the cells were classified as infected. The threshold was defined by maximizing the difference in the infection rates between the positive and negative controls, which were spotted on the same plate. Quality filtering was performed to eliminate the out-of-focus images and image artifacts. On the single image level, the images were automatically classified as low quality if they contained too few or too many cells or if they were out-of-focus. On the whole plate level, the percentage of saturated pixels in channel 2 was computed. Over-exposed plates were scanned again using decreased exposure times [92].

2.1.4 Clustering of infected cells

K-function

The K-function or Ripley's K-function is a well-established measure for defining the degree of clustering. This measure evaluates all interparticle distances over the studied area and compares the observed distribution with a random distribution of spots. Ripley's K-function has been used in ecology, epidemiology and geography [41]. In cell biology, the function was applied to study the integrin-sensing extracellular matrix properties [102] and to analyze lipid rafts by observing the clustering of the RAS proteins [110].

The distribution of cells in fluorescence microscopy images was represented as a spatial pattern of spots. The spots (cells) were classified as infected or non-infected, and their respective clustering behaviors were studied using the K-function. The K-function function was introduced by Ripley in 1977 [118]. The K-function is a distance-based method of measuring the ratio of the expected number of neighbors within a circle with a given radius $c(r)$ to the expected density. The K-function is calculated using the equation

$$K(r) = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \frac{1}{w(x_i, d_{ij})} \frac{I_r(d_{ij})}{N} \tag{2.1}$$

for a given radius parameter $r>0$. The variables of the equation include the following: N is the number of spots in the observed area A (whole image); λ is the intensity of spots, which can be estimated by $\frac{N}{A}$; d_{ij} is the Euclidean distance between the spots i and j. The function $I_r(d_{ij})$ is equal to one if $d_{ij} < r$ and is zero otherwise. Since each point's neighborhood is only defined within a given study area, points close to the area's boundary need to be processed with edge correction to get an accurate estimation. The weighting factor $w(x_i, d_{ij})$ corrects for the edge effects and is the proportion of the circumference of a circle with center x_i and distance d_{ij} that falls in the studied area. Let $c_i^+(r)$ and $c_i^-(r)$ be the regions of the search circle of a point i that belong or do not belong to the study area, respectively. We usually do not know the number of points within $c_i^-(r)$. If the points in this area are not considered, we might find points in $c_i(r)$ that are lower than expected. Suppose that the point density within $c_i^-(r)$ is equal to the point density within $c_i^+(r)$. Let us define the total number of points within $c_i(r)$ as the following:

$$n_i(r) = n_i^+(r) + n_i^-(r) \tag{2.2}$$

where $n_i^+(r)$ and $n_i^-(r)$ are the number of points within $c_i^+(r)$ and $c_i^-(r)$, respectively. The area of the circle, $c_i(r)$, can be defined as

$$Area_i(r) = Area_i^+(r) + Area_i^-(r) = \pi r^2 \tag{2.3}$$

where $Area_i^+(r)$ and $Area_i^-(r)$ are the areas of the region within $c_i^+(r)$ and $c_i^-(r)$, respectively. Using the density definition and above assumption, we find the following:

$$\begin{aligned} \frac{n_i^-(r)}{Area_i^-(r)} &= \frac{n_i^+(r)}{Area_i^+(r)} \\ n_i^-(r) &= \frac{Area_i^-(r)}{Area_i^+(r)} n_i^+(r). \end{aligned} \tag{2.4}$$

From equation (2.2), (2.3) and (2.4), it follows that

$$n_i(r) = \frac{Area_i(r)}{Area_i^+(r)} n_i^+(r) = \frac{\pi r^2}{Area_i^+(r)} n_i^+(r) = \frac{1}{w(x_i, d_{ij})} n_i^+(r). \tag{2.5}$$

When the circle is entirely inside of the studied area, $Area_i^+(r)$ is equal to πr^2, $n_i(r) = n_i^+(r)$ and $w(x_i, d_{ij}) = 1$. From equation (2.1) and equation (2.5), we obtain the following:

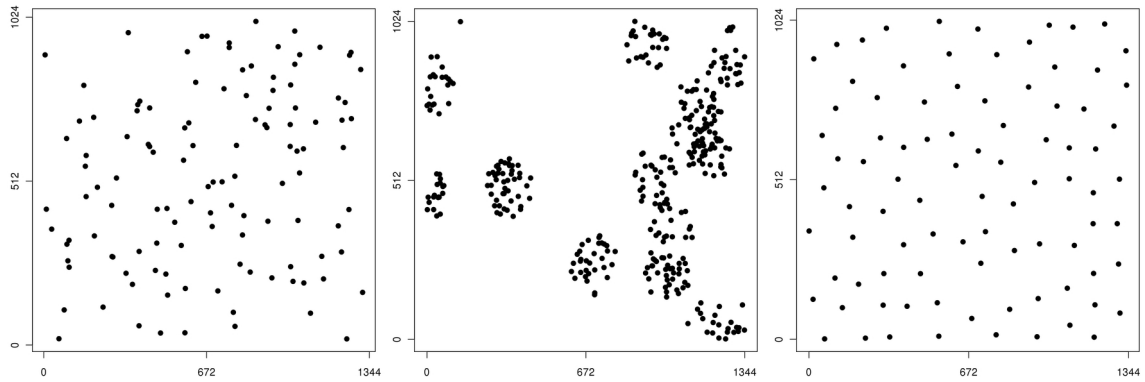

Figure 2.2: Examples of three different point patterns. (a) random distribution of spots; (b) clustering spots; (c) regular pattern. The normalized clustering scores for the random distribution, clustering spots, and regular patterns are 0.03, 0.98, and -1.02, respectively.

$$\hat{K}(r) = \frac{1}{\lambda}\frac{1}{N}\sum_{i=1}^{N} n_i(r) = \frac{1}{\lambda}\pi r^2 \left(\frac{1}{N}\sum_{i=1}^{N}\frac{n_i^+(r)}{Area_i^+(r)}\right) = \frac{1}{\lambda}\cdot\left(\hat{\lambda}\pi r^2\right), \tag{2.6}$$

where the $\hat{\lambda}$ is the density estimation. Ripley's K-function is used to compare the observed spot distribution with a random distribution. The given spot distribution is tested against the null hypothesis that the spots are randomly distributed. For clustering distributions, the expected value of *K(r)* is larger than the value of a random distribution; for regular patterns, this expected value is less than for a random distribution. For the complete spatial randomness (CSR) and assuming the points are randomly distributed, the average neighborhood density is equal to πr^2. In equation (2.6), if the points adhere to the CSR, $\hat{\lambda} = \lambda$ and $\hat{K}(r) = \pi r^2$. Furthermore, if $\hat{\lambda} < \lambda$, then the average neighborhood density is less than the expected, which means that points are dispersed and the $\hat{K}(r) < \pi r^2$. If $\hat{\lambda} > \lambda$, then the average neighborhood density is greater than expected, which means that the points are clustered and $\hat{K}(r) > \pi r^2$ [130]. Figure 2.2 shows examples of the spot distributions for the random, clustering, and regular distribution; the plot of $\hat{K}(r)$ is shown in Figure 2.3.

To correct for the biases caused by the clustering of proliferating cells, we used a

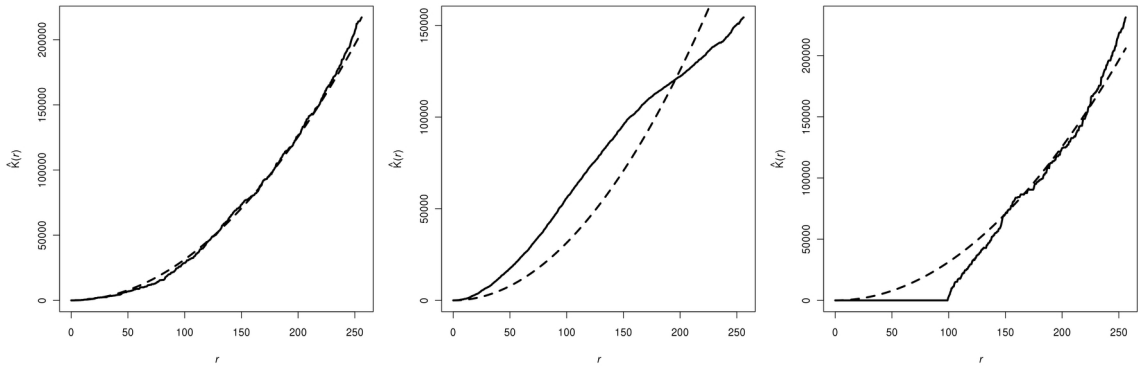

Figure 2.3: Estimated K-values for the point patterns. Solid curves are the plots of the inhomogeneous K-function for the data of Figure 2.2. (a) random distribution of spots; (b) clustering spots; (c) regular pattern). Dashed curves represent the K values of the complete spatial randomness (CSR) distributions, which were used as controls. The normalized clustering scores for the random distribution, clustering spots, and regular patterns are 0.03, 0.98, and -1.02, respectively.

random distribution that used the actual positions of the spots of infected and non-infected cells. The s^{th} simulated null-hypothesis of the K-function was estimated by randomly drawing N_c spots from all spots (infected and non-infected cells) and applying them to the K-function. The final null-hypothesis was calculated from the mean value of these simulated K-functions (s = 1, ...,100). We applied K-function to the spot distributions using the local spatial variation (independent from their clustering) and the inhomogeneous K-function as defined by Baddeley and co-workers [3]. The inhomogeneous K-function is given by the following equation:

$$K_{\text{inhom}}(r) = \frac{1}{|A|} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \frac{e_{ij} I_r(d_{ij})}{\lambda(y_i)\lambda(y_j)} \tag{2.7}$$

where $|A|$ denotes the observation area (distance $\leq$ r) and e_{ij} is the edge-correction factor calculated by the border method [119]. $\lambda(y_i)$ and $\lambda(y_j)$ are estimated intensities at spots y_i and y_j. These variables were estimated using a Gaussian kernel smoother and the intensity surface model [3]. The maximum ranges of the radius r that we investigated were 25%, 30%, 35%, and 40% of the shorter side of the whole image. To obtain the clustering score, the area between the curves of the

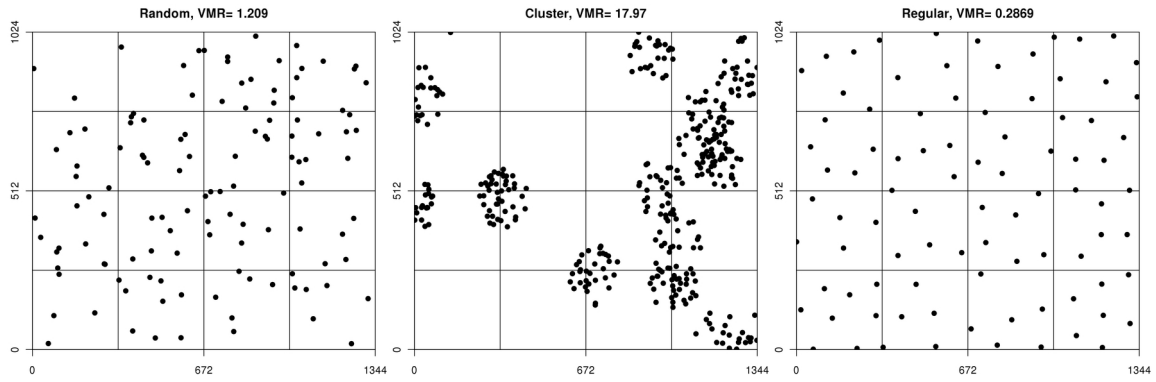

Figure 2.4: Three point pattern distributions with different *VMR* scores. The *VMR* scores for the random distribution, clustering spots, and regular patterns are 1.21, 17.97, and 0.27, respectively.

inhomogeneous K-function and a simulated random distribution was calculated. The score was positive if the curve for the inhomogeneous K-function was mainly above the curve of the simulated random distribution (tendency for clustering), and the score was negative otherwise. This score was calculated for the infected and non-infected cells, respectively. To estimate the infection rate using the final clustering score, the score of the infected cells was subtracted by the score of the non-infected cells. The library spatstat [2] in the R-programming environments was used to compute the estimated K-value.

Quadrat Analysis

We also observed the clustering of cells with another clustering method, which is called a quadrat analysis. The quadrat analysis observes the frequency distribution of cells within a set of grid squares (quadrat) [156]. To obtain the variance-mean ratio (VMR) as a measure of the clustering of points, the mean number of cells per quadrat is estimated, and its variance is computed using the following:

$$VMR = \frac{s^2}{\bar{x}}, \tag{2.8}$$

$$s^2 = \sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{m-1}, \tag{2.9}$$

where m is the number of quadrats, x_i is the number of points in quadrat i and $\bar{x}$ is the mean of the number of points per quadrat. A *VMR* value of greater than one indicates a clustered distribution, a *VMR* value of less than one indicates a random distribution and a $VMR = 0$ indicates a uniform distribution. The *VMR* scores for the random, cluster and regular distributions are shown in Figure 2.4. The *VMR* scores were computed with a 4x4 grid quadrant and yielded the following: a score of 1.209, which is nearly one for the random distribution; a score of 17.97, which is much higher than one for the clustered distribution; and a score of 0.29, which is less than one for the uniform or regular distribution. To obtain the final clustering score, we subtracted the *VMR* scores of the non-infected cells from the *VMR* scores of the infected cells. The clustering score was calculated for all knocked-down genes and the controls, and a z-normalization was performed.

2.1.5 Comparing the clustering results and experimental results

We investigated the results of the clustering approaches by comparing the z-scores from K-function and Quadrat Analysis for all knocked-down genes with the z-score from the intensity readouts of the primary and secondary screens. The Pearson correlation coefficient was employed for this comparison. Given the scores of all knocked-down genes from the clustering approach, $X = (x_1, x_2, x_3, \ldots, x_n)$, and the intensity readouts from the experiments, $Y = (y_1, y_2, y_3, \ldots, y_n)$. The correlation coefficient can be computed as follows:

$$R = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} \tag{2.10}$$

where $cov(X, Y)$ is the covariance between X and Y, $cov(X, Y) = \sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})$, and var is the variance of the data, $var(X) = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $var(Y) = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$. The correlation coefficient values range from -1 to +1. If the correlation coefficient is close to 1, the clustering scores are consistent with the intensity values from the experiments, whereas a correlation coefficient that is close to -1 suggests that the score and the intensity have an opposite tendency. A correlation coefficient of 0 means that no linear relationship exists between the score and the intensity.

2.1.6 The statistical method used to identify the host siRNA hits

For the primary screen, we excluded the wells with less than 125 and more than 500 cells. For the secondary screen, the wells showing the lowest and highest 5% of the firefly reporter activity (correlated to the number of viable cells) were excluded. These wells were excluded to eliminate possible interference with the readout of viral replication from cytostatic or cytotoxic effects or high variability in the cell number. In some wells, the cells might have grown densely and it is possible that incorrect segmentation of images occurred [9]. The viral-specific signal intensities per siRNA were normalized for the effects of differing cell counts using a local-weighted scatterplot smoothing method [25].

The B-score normalization was used to remove the spatial effects within individual LabTeks [13] and accounted for the row and column variation effects. The variability between plates was addressed by subtracting the plate median from each measurement per siRNA and then dividing the resulting value by the plate median absolute deviation (1σ), which resulted in one score per siRNA per LabTek. The advantage of the B-score normalization is that it minimizes the biases due to positional effects [89]. To compute the B-score, first we calculated the residual r_{ijp} for the row i and column j on the plate p which is defined as the following:

$$r_{ijp} = y_{ijp} - \hat{y}_{ijp} = y_{ijp} - (\hat{\mu}_p + \hat{R}_{ip} + \hat{C}_{jp}). \tag{2.11}$$

The residual is the difference between the measured value y_{ijp} and the fitted value $\hat{y}_{ijp}$ that is computed from the estimated average of the plate ($\hat{\mu}_p$) and the estimated systematic measurement offsets for each row i on plate p ($\hat{R}_{jp}$) and column j on plate p ($\hat{C}_{jp}$). The B-score is calculated by the following:

$$\text{B-Score} = \frac{r_{ijp}}{\text{MAD}_p}, \tag{2.12}$$

where MAD_p is the adjusted median absolute deviation for each plate p; MAD_p is a robust estimate of the spreading of r_{ijp}: $\text{MAD}_p = median\,\{|\, r_{ijp} - median(r_{ijp})|\}$. The replicates were summarized using the mean of the normalized scores; furthermore, Student's t-tests were carried out to determine whether the siRNA effects differed significantly from zero. Only the hits with negative z-scores were taken.

For all three analyses (the primary screen, secondary screen, and clustering analysis), the hits were selected if their P-values were below 0.05. The statistical analysis of the processed imaging data was carried out using the R-programming language and integrating the Bioconductor libraries RNAither [117] and cellHTS [12].

2.2 Characterization of the signaling interactions

2.2.1 General concept and workflow

An overview of our workflow is shown in Figure 2.5. First, the cellular phenotypes from the RNAi screening images were quantitatively measured and analyzed. Protein interactions (activation and inhibition of signal transduction) were assembled from the database KEGG [66, 67, 155], and these interactions were used as a gold standard. We generated novel phenotypic features describing the similarity of phenotypes between two proteins. In addition, the phenotypic features from the original study of the image data [98] were assembled. We then established a systematic classification using the Support Vector Machines (SVMs) that was based on the phenotypic descriptors used to classify the set of interactions that activate or inhibit signal transduction. The trained machines were evaluated and then used as a prediction model for unknown interactions. All interactions were used to define a similarity score, which is called the consistency score. The performance was improved using this score. The consistency score was verified with other interaction databases. We applied the consistency score for a detailed analysis of the cytokine receptor signaling. Unsupervised clustering was performed to find proteins that have similar functions. A cluster with a predicted domain of interaction was investigated in further detail.

2.2.2 Data sources

List of interactions from KEGG that activate or inhibit signal transduction

The characterizations of the PPIs in signaling pathways that were used to construct the human signaling network were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG, www.genome.jp/kegg) [66, 67, 155]. The KEGG provides a comprehensive set of interactions which are linked to the supporting literature

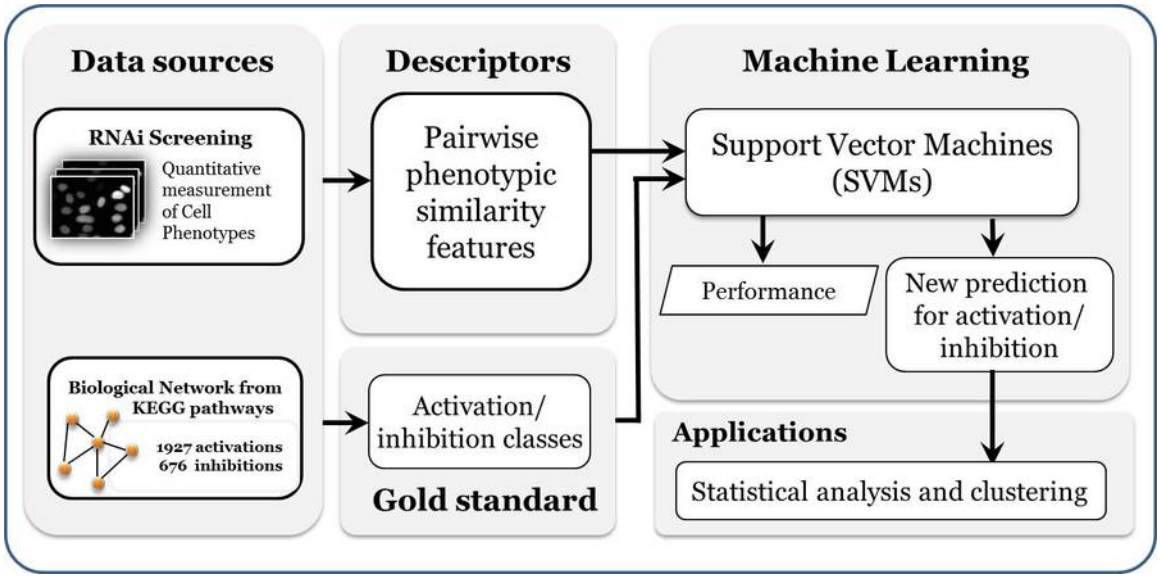

Figure 2.5: The workflow for the prediction of interactions that are involved in the activation and inhibition of signal transduction using the pairwise phenotypic similarity features and SVMs.

evidence. We used the lists of activation (Act-PPI) and inhibition (Inh-PPI) from eleven signaling pathways (Table 2.1) that had a high overlap with the cytokine receptors such as those from the endocrine signaling system, cell growth and death and the immune system. In total, we had phenotypic data for 663 proteins for which we had phenotypic data were investigated. Among these, we got 1927 known activation and 676 known inhibition interactions. The protein pairs of all sets (Act-PPI, Inh-PPI) were further analyzed.

We also used the PPIs from the Search Tool for the Retrieval of Interacting Genes/proteins (STRING) version 9.0 [136] and the MetaCore^{TM} (www.genego.com) to perform the enrichment analysis. The STRING database includes an interaction database of known and predicted PPIs. The MetaCore database is an interaction database that provides additional pairs of interacting proteins.

Cellular imaging data

The morphological changes in the nuclei of HeLa cell clones that were stably transfected with the GFP-tagged histone 2B were tracked by fluorescence imaging after the transient transfection of the siRNAs in the high-throughput screens. The

Table 2.1: The list of the selected pathways from the KEGG database.

Signaling pathways	**KEGG ids**	**Pathway groups**
Insulin	hsa:04910	Endocrine system
VEGF	hsa:04370	Signal transduction
MAPK	hsa:04010	Signal transduction
ERBB	hsa:04012	Signal transduction
mTOR	hsa:04150	Signal transduction
WNT	hsa:04310	Signal transduction
TGF-beta	hsa:04350	Signal transduction
Jak-STAT	hsa:04630	Signal transduction
Cell cycle	hsa:04110	Cell Growth and Death
Chemokine signaling pathway	hsa:04062	Immune System
Cytokine-cytokine receptor interaction	hsa:04060	Signaling molecules and interactions

cells were distributed on the cell microarrays (labteks) that were printed with the transfection-ready siRNAs, and the chromosome/nuclear morphology was visualized in real-time. One image contained more than 100 nuclei with an average diameter of approximately 30 pixels in the G1 phase. All images had a grey value depth of 16 bit and a spatial resolution of 1344x1024 pixels. Each image sequence consisted of 96 time points over 48 hours. The analyzed images were obtained from the Mitocheck Database.

2.2.3 Machine learning for classification: the LDA and SVM

The machine learning approach is a computational method for the design and development of algorithms capable of learning empirical data. A major task of the machine learning approach is to recognize patterns and then makes an intelligent decision based on the learned data. The machine learning approach is mainly categorized into supervised and unsupervised learning methods. A supervised learning method requires the training data with correctly predefined classes for learning and produces an inferred function (a classifier or a regression function). In contrast, an unsupervised learning approach, such as clustering and association rules, is based

on a data distribution in a feature space without predefined classes and describes hidden patterns in the data. The general elements of the classification task in the machine learning approach are comprised of the following: 1) learned data, 2) learning algorithms (*e.g.*, the Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes), and 3) performance evaluations of the classifier. To measure the ability of the classifier to perform accurately on new (untrained) data, the learned data are divided into training and testing sets. The learning algorithm learns from the training set and tests on the testing set.

In this project, we focused on the supervised learning approaches, such as linear discriminant analysis (LDA) and a Support Vector Machine (SVM). We used LDA and a SVM for several tasks. LDA was trained to classify two sets of cellular phenotypes resulting from the knockdown of different genes. The accuracy of this classification was the similarity of the two knockdown phenotypes. For the SVM, we first used it to classify each single cell into four phenotypic classes (apoptosis, interphase, mitosis, and shape) and computed the fraction of each phenotype with respect to the number of cells in an knockdown image. Second, we used the SVM to distinguish the activities of the PPIs, which consisted of the activation and inhibition of signal transduction. Both LDA and the SVM are supervised machine learning approaches. The supervised learning method requires prior knowledge of a set of objects, which are composed of values of their descriptors and class labels. For the LDA, the descriptors are the image features (Section 2.2.4), and the classes are the two groups to which the single cells belong. For the SVM classification of the four phenotypes, the descriptors are also the image features (Section 2.2.4), and the classes are the four class labels. For the SVM classification of the types of PPIs, the descriptors are the pairwise phenotypic descriptors (Section 2.2.5), and the classes are the labels of activation and inhibition. After the training procedure, the classifiers are applied to superimpose the class labels from the given descriptors on new objects for which the class labels are unknown. The principle of LDA and the SVM is briefly described in the following sections.

2.2.3.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is an approach widely used in classifications that are based on the linear combinations of feature vectors. The method performs feature dimensionality reduction while preserving the class separability and charac-

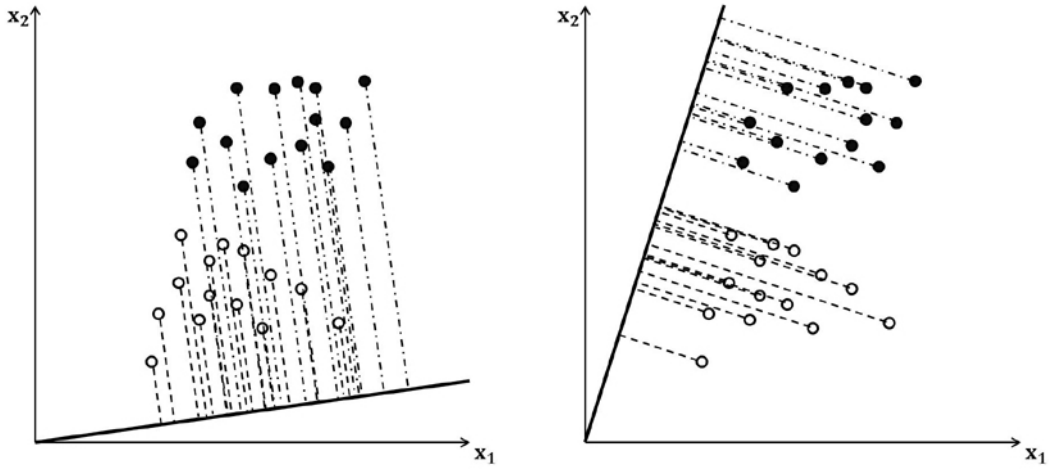

Figure 2.6: Two-dimensional case of projecting the sample on a line. Figure (a) shows two sample groups that are mixed on the projected line. Figure (b) shows two sample groups that are separated by the projected line.

terizes two or more classes of the data with the highest conditional probability. The resulting combination uses a linear classifier. A criterion of the linear discriminant is that the ratio of the between-class to within-class scatter must be maximized. The sample $\mathbf{x}$ is projected onto a line by $\mathbf{y} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$. The optimal line is the line that maximizes the separation of two or more classes. Figure 2.6(b) shows the optimal line for the separation of the two-dimensional samples, whereas the two groups cannot be separated by the line in Figure 2.6(a).

To find a suitable projection vector, the mean vector of each class in the $\mathbf{x}$ and $\mathbf{y}$ feature spaces uses the following as a measure:

$$\mu_i = \frac{1}{N_i} \sum_{\mathbf{x} \in c_i} \mathbf{x}, \tag{2.13}$$

$$\text{and } \tilde{\mu}_i = \tfrac{1}{N_i} \textstyle\sum_{\mathbf{y} \in c_i} \mathbf{y} = \tfrac{1}{N_i} \sum_{\mathbf{x} \in c_i} \mathbf{w}^{\mathrm{T}}\mathbf{x} = \mathbf{w}^{\mathrm{T}}\mu_i.$$

The objective function is the distance between the projected means and is given by the following:

$$J(\mathbf{w}) = |\tilde{\mu}_1 - \tilde{\mu}_2| = \left|\mathbf{w}^{\mathrm{T}}(\mu_1 - \mu_2)\right|. \tag{2.14}$$

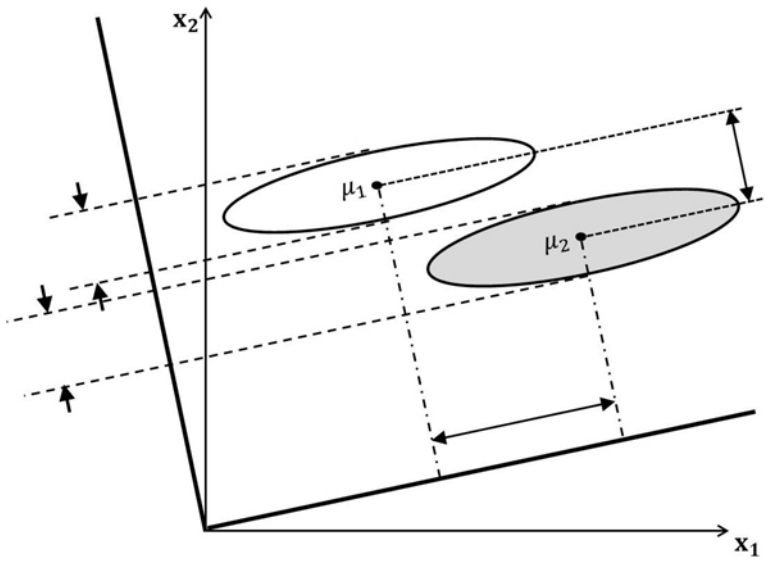

Figure 2.7: An example of a data projection considering the means and standard deviations.

However, the consideration of the mean alone is not enough. The standard deviation within the classes should also be taken into account because in some cases the difference between the means is high, but the data of each group are scattered and highly overlapping (Figure 2.7). Fisher [43] proposed a solution to this problem that maximize the difference between the means normalized by a measure of the within-class scatter. The variance of each class can be defined as the following:

$$\tilde{\mathbf{s}}_i^2 = \sum_{\mathbf{y} \in c_i} (\mathbf{y} - \tilde{\mu}_i)^2, \tag{2.15}$$

where the quantity $(\tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2)$ is called the within-class scatter of the projected data. The linear discriminant is defined as the linear function $\mathbf{w}^{\mathrm{T}}\mathbf{x}$ that maximizes the following criterion function:

$$J(\mathbf{w}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2}. \tag{2.16}$$

We find a projection where the samples in the same class are close together and the projected means are the furthest apart from one another. To find the optimal

projection $\mathbf{w}$, the explicit form of $\mathbf{w}$ from the $J(\mathbf{w})$ needs to be expressed. In the scatter of the multivariable feature space $\mathbf{x}$, the scatter measurements are given by the following scatter matrices:

$$\mathbf{S}_i = \sum_{\mathbf{x} \in c_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^{\mathrm{T}}. \tag{2.17}$$

In the two-class classification, we defined the within-class scatter matrix $\mathbf{S_w}$ where the $\mathbf{S_w} = \mathbf{S}_1 + \mathbf{S}_2$. The scatter of the projection $\mathbf{y}$ can then be determined in a function of the scatter matrix in the feature space $\mathbf{x}$:

$$\tilde{\mathbf{s}}_i^2 = \sum_{\mathbf{y} \in c_i} (\mathbf{y} - \tilde{\mu}_i)^2 = \sum_{\mathbf{x} \in c_i} (\mathbf{w}^{\mathrm{T}}\mathbf{x} - \mathbf{w}^{\mathrm{T}}\mu_i)^2 = \sum_{\mathbf{x} \in c_i} \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^{\mathrm{T}}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{S}_i\mathbf{w}. \tag{2.18}$$

Therefore, we obtain the following:

$$\tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2 = \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}. \tag{2.19}$$

Similarly, the difference between the projected means can be expressed in terms of the means in the feature space $\mathbf{x}$ as given by the following:

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (\mathbf{w}^{\mathrm{T}}\mu_1 - \mathbf{w}^{\mathrm{T}}\mu_2)^2 = \mathbf{w}^{\mathrm{T}}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}, \tag{2.20}$$

where $\mathbf{S}_{\mathrm{B}}$ is denoted as the between-class scatter. Notably, the rank of $\mathbf{S}_{\mathrm{B}}$ is at most one because it is the outer product of two vectors. By substituting equation (2.19) and equation (2.20) into equation (2.16), we obtain the Fisher criterion as the following:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}. \tag{2.21}$$

To find the maximum of $J(\mathbf{w})$, we compute the derivative of $J(\mathbf{w})$ and set it equal to zero:

$$\begin{gathered}
\frac{d}{d\mathbf{w}} J(\mathbf{w}) = \frac{d}{d\mathbf{w}} \left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \right) = 0, \\
\left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}\right) \frac{d}{d\mathbf{w}} \left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}\right) - \left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}\right) \frac{d}{d\mathbf{w}} \left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}\right) = 0, \\
\left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}\right) (2\mathbf{S}_{\mathrm{B}}\mathbf{w}) - \left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}\right) (2\mathbf{S}_{\mathrm{W}}\mathbf{w}) = 0, \\
\left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \right) (\mathbf{S}_{\mathrm{B}}\mathbf{w}) - \left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \right) (\mathbf{S}_{\mathrm{W}}\mathbf{w}) = 0, \\
\mathbf{S}_{\mathrm{B}}\mathbf{w} - \lambda\mathbf{S}_{\mathrm{W}}\mathbf{w} = 0, \text{ where } \lambda \text{ is a constant},
\end{gathered} \tag{2.22}$$

$$\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}\mathbf{w} = \lambda\mathbf{w}. \tag{2.23}$$

In this case, it is unnecessary to solve for the eigenvalues and eigenvectors of $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ because $\mathbf{S}_{\mathrm{B}}\mathbf{w} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{w} = (\mu_1 - \mu_2) \cdot k$, where k is a constant, is always in the direction of $\mu_1 - \mu_2$ and the scale factor for $\mathbf{w}$ is unimportant [35]. Therefore, the unscaled solution for the $\mathbf{w}$ that optimizes $J(\cdot)$ is $\mathbf{w} = \mathbf{S}_{\mathrm{W}}^{-1}(\mu_1 - \mu_2)$. Thus we have obtained $\mathbf{w}$ for Fisher's linear discriminant, which is the linear function that produces the maximum ratio of between-class scatter to within-class scatter. The classification has been converted from a d-dimensional problem to a one-dimensional problem. We then find the threshold that is the point along the one-dimensional subspace separating the projected points. The optimal decision boundary has the equation $\mathbf{w}\mathbf{x} + w_0 = 0$ where $\mathbf{w} = \mathbf{S}_{\mathrm{W}}^{-1}(\mu_1 - \mu_2)$ and w_0 is a constant involving $\mathbf{w}$ and the prior probabilities. The optimal decision rule is to decide data in c_1 if the linear discriminant exceeds some threshold, and to decide c_2 otherwise.

2.2.3.2 Support Vector Machines

In the field of pattern recognition, the Support Vector Machines (SVMs) [16, 146] have been widely used for classification purposes. SVMs are effective supervised learning algorithms for finding an optimal hyperplane that separates the sample classes of training data by maximizing the distance to the nearest training data points. In the following, we briefly describe the basic concepts of SVMs.

Linear Support Vector Machine

We consider the linear separable binary classification or the **separable case**. For a given l training samples with a dimensionality D, $\{x_i, y_i\}$, $i = 1, ..., l$ where $x_i \in \Re^D$ and $y_i \in \{-1, 1\}$ are the respective classes, and we assume that the samples are linearly separable. The separating hyperplane is defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is the normal vector of the hyperplane and $b/\left\|\mathbf{w}\right\|$ is the perpendicular distance from the hyperplane to the origin. The support vectors are the data points closest to the separating hyperplane and defined by the margin. The margin is given by the two parallel hyperplanes $\mathbf{H}_1, \mathbf{H}_2$ with equal distance to the separating hyperplane (Figure 2.8(a)). The aim of the SVM is to orientate this hyperplane to be furthest from the closest samples of both classes, which maximizes the margin. Suppose that all the training data satisfy the following constraints:

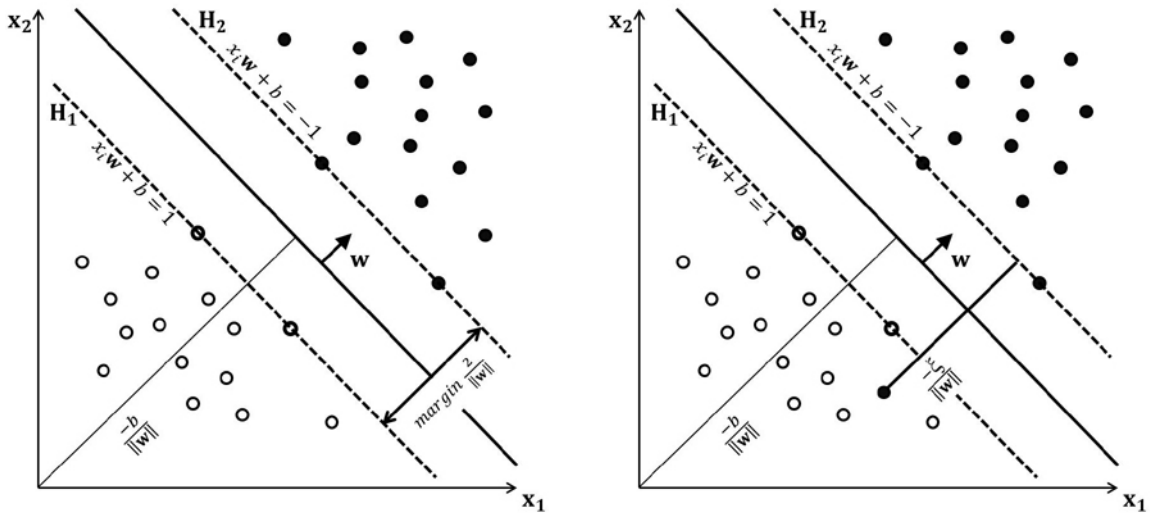

Figure 2.8: Linear separation in a two-dimensional feature space. (a) The SVM attempts to find an optimal linear hyperplane by maximizing the margin. The dashed lines are the margins chosen with the closest data points to the line. The data points that constrain the width of the margins are called the support vectors. For the (b) non-separable case of the SVM, the constraint is relaxed by the introduction of a slack variable.

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \text{ for } y_i = +1 \tag{2.24}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \text{ for } y_i = -1. \tag{2.25}$$

These inequalities can then be combined into the following:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0, \ \forall \ i \text{ and } y_i \in \{-1, 1\}. \tag{2.26}$$

The support vectors are then the data points lying on the following two hyperplanes:

$$\begin{aligned} \mathbf{H}_1 &: \mathbf{x}_i \cdot \mathbf{w} + b = 1 \\ \mathbf{H}_2 &: \mathbf{x}_i \cdot \mathbf{w} + b = -1 \end{aligned} \tag{2.27}$$

which the margin is defined as the distance between these two hyperplanes. We calculate the margin by subtracting the perpendicular distance of $\mathbf{H}_2$ to the origin

($|-b+1|/\|\mathbf{w}\|$) from the perpendicular distance of $\mathbf{H}_1$ to the origin ($|-b-1|/\|\mathbf{w}\|$). Hence, the margin is simply $2/\|\mathbf{w}\|$. Thus, we find the pair of hyperplanes that give the maximum margin by minimizing $\|\mathbf{w}\|$ subject to the constraints of equation (2.26). To avoid the square root in the norm and allow Quadratic Programming (QP) to be used later on, we minimize $\frac{1}{2}\|\mathbf{w}\|^2$, which is equivalent to minimizing $\|\mathbf{w}\|$. We therefore need to solve the following:

$$\begin{aligned} &\min && \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{subject to} && y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \ \geq 0, \ \forall\, i. \end{aligned} \tag{2.28}$$

The method of Lagrange multipliers can be used to find the minima of this objective function subject to the constraint. The Lagrange multipliers, $\alpha_i \geq 0$, $i = 1, \ldots, l$, are introduced. The Lagrangian is the following:

$$\begin{aligned} L_P &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1\right] \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{w} - b\sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i, \ \ \alpha_i \geq 0, \, \forall i. \end{aligned} \tag{2.29}$$

We then compute the partial derivatives $\frac{\partial}{\partial \mathbf{w}} L_p$ and $\frac{\partial}{\partial b} L_p$ and set them equal to zero:

$$\frac{\partial}{\partial \mathbf{w}} L_P = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \tag{2.30}$$

$$\frac{\partial}{\partial b} L_P = 0 \quad \Rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.31}$$

By substituting equation (2.30) and equation (2.31) into equation (2.29), we find the following:

$$\begin{aligned} L_D \ &= \ \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j, \\ &\text{subject to} \ \ \alpha_i \geq 0, \ \forall i \ \text{ and } \ \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned} \tag{2.32}$$

This L_D is referred to as the dual form of the primary L_P. Note that L_D only depends on the Lagrange multiplier α (not on $\mathbf{w}$ and b); in L_D, the training data

appear as the dot products ($\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$), and this property can be exploited to perform the classification in a higher dimensional space. For the training of the SVM, we maximize L_D with respect to α_i and subject to the constraints of equation (2.31) and the positivity of the α as shown in equation (2.32). $\mathbf{w}$ is then given by equation (2.30), and b is determined with the *complementary condition* of the *Karush-Kuhn-Tucker* (KKT) conditions for the primal problem L_P:

$$\alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \right] = 0, \ \forall i. \tag{2.33}$$

The complementary condition of the KKT is applied to all samples in the training set. Therefore, for each sample, either $\alpha_i = 0$ or $(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1) = 0$ must be hold. Those sample points for which $\alpha_i > 0$ are then found on one of the hyperplanes which are the support vectors. For all other training samples $\alpha_i = 0$.
The sample point, which is a support vector ($\mathbf{x}_s$), will have the following form:

$$y_s(\mathbf{x}_s \cdot \mathbf{w} + b) = 1. \tag{2.34}$$

We substitute equation (2.30) into equation (2.34) and find the following:

$$y_s(\sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s + b) = 1 \tag{2.35}$$

where S denotes the set of indices of the support vectors. S is determined by finding the indices i where $\alpha_i > 0$. We then multiply through by y_s when $y_s^2 = 1$. Therefore, we get b from the following:

$$y_s^2(\sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s + b) = y_s \quad \Rightarrow \quad b = y_s - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s. \tag{2.36}$$

Instead of using an arbitrary support vector $\mathbf{x}_s$, it is more advantageous to take an average of all of the support vectors in S:

$$b = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s). \tag{2.37}$$

To apply the trained SVM for the classification of a test sample $\mathbf{x}_t$, we applied the following hyperplane decision function:

$$f(\mathbf{x}_t) = sign(\mathbf{w} \cdot \mathbf{x}_t + b). \tag{2.38}$$

For the binary classification of the data that is not fully linearly separable or the **non-separable case**, the constraints of equation (2.26) are relaxed to allow for misclassification of the samples by the introduction of a positive slack variable, ξ_i, $i = 1, \ldots, l$ (Figure 2.8(b)) as follows:

$$\begin{aligned} y_i(\mathbf{x}_i \cdot \mathbf{w} + b) &\geq 1 - \xi_i, \, \forall i, \\ \xi_i &\geq 0, \, \forall i. \end{aligned} \tag{2.39}$$

These slack variables measure the deviation from the ideal conditions. For $0 \leq \xi_i \leq 1$, the data point falls inside the region of separation but on the right side of the decision surface. For $\xi_i > 1$, the data point falls on the wrong side of the separating hyperplane. The sum of the slack variables $\sum_i \xi_i$ provides an upper bound on the number of training errors. The objective function can be formulated in the relaxed version as $\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i$, where the parameter C regulates the penalty of errors and has to be chosen by the user. This formulation is called a *soft margin* classifier. Formulating the primal problem by applying the Lagrange multipliers α_i and μ_i yields the following:

$$L_P = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i - \sum_i \alpha_i \{\mathbf{y}_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i \tag{2.40}$$

where μ_i is introduced to enforce the inequality $\xi_i \geq 0$. Differentiating with respect to $\mathbf{w}$, b and ξ_i and setting the derivatives to zero:

$$\frac{\partial}{\partial w} L_P = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \tag{2.41}$$

$$\frac{\partial}{\partial b} L_P = 0 \quad \Rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \tag{2.42}$$

$$\frac{\partial}{\partial \xi_i} L_P = 0 \quad \Rightarrow \quad C = \alpha_i + \mu_i. \tag{2.43}$$

The substitution of these formulations into equation (2.40) gives the formulation of the dual problem L_D, which is the same for the separable case. The only difference is that there is an additional constraint $0 \leq \alpha_i \leq C$, $\forall i$, which means that there exists an upper bound C on the α_i in this non-separable case. b is then calculated in the same way as the separable case with the KKT conditions of the primal problem:

$$\alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \right] = 0 \text{ and } \mu_i \xi_i = 0. \tag{2.44}$$

Non-linear Support Vector Machine
The generalization of the above formulations with the non-linear decision function is straightforward. The training samples are mapped to a higher dimensional Euclidean space H by a non-linear feature mapping function Φ. The mapping is performed in accordance with Cover's theorem, which the data in mapped space are linearly separable. For the training of the SVM equation (2.32), the training data only appear in a dot product $\mathbf{x}_i \cdot \mathbf{x}_j$. Thus, for the data transformed to H, the machine handles only the dot product of the mapping $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. If there exists a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, we do not have to consider the explicit form of Φ, but could only use $K(\mathbf{x}_i, \mathbf{x}_j)$ instead of using the dot product $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ in equation (2.32). Thus, the SVM performs a linear separation of the data in H corresponding to a non-linear separation in the lower dimensional original space with the following:

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{2.45}$$

There exist kernel functions with the property $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ if they satisfy *Mercer*'s condition: for any $g(\mathbf{x})$ such that $\int g(\mathbf{x})^2 d\mathbf{x}$ is finite, then $\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0$. This condition only examines whether a kernel is an inner-product kernel in some space, but it does not tell us how to construct the mapping function Φ. We used the Gaussian radial basis function kernel in our analysis because it has been shown to work very well for the classification of cell images in previous analyses [27, 55].

2.2.4 Image features for the classification of cells

To analyze images of the siRNA screens, an automated system, which was described in detail recently [55, 56, 57], was employed. Briefly, a quadratic sliding window was used to calculate local thresholds for different image regions. The local threshold was only calculated if the variance within the window reached a pre-defined threshold (2000); otherwise, a global threshold was used. The window consisted of an outer

region of 15 pixels in which the thresholds were computed and an inner region of 2 pixels in which the thresholds were applied. The window was shifted by the length of its inner region. The global and local thresholds were calculated using the Otsu thresholding method. After segmentation, the following quantitative image features were extracted from the image for each single cell: granularity features, object- and edge-related features, tree-structured wavelet features, Haralick texture features, grey-scale invariants and Zernike moments. In total, we computed 353 features for each cell nucleus. Using these features, the single-cell images were classified into the following classes: interphase, mitosis, apoptosis and cell clusters. The Haralick features have relatively high computational costs. It would have taken several weeks or months to compute the Haralick texture features from all the image data; therefore, my colleagues, M.Gipp, G.Marcus and R.Männer, and us employed the general-purpose graphics processing units (GPUs) to speed up the computation of the co-occurrence matrices and Haralick texture features. A massive parallel software version for the GPUs was designed and implemented for this purpose. The computational time was shortened by a factor of 32 on a single node of a cluster in comparison to a pre-existing optimized CPU software version [47].

We extracted a set of image features for each single cell. Table 2.2 illustrates the number of extracted features. All these features were described in detail in Harder *et al.* [55, 56, 57]; I briefly describe them below.

Haralick texture features: The Haralick texture features [53] are the most important features and have been widely used to describe the characteristics of a cell image in several research reports [28, 45, 55, 98]; these features are also included in several cell analysis software packages [29, 64, 103]. The Haralick texture features are based on the co-occurrence matrices of an image. A co-occurrence matrix (C) is computed by the relative frequencies of all occurring gray value pairs of pixels at a given distance d with the angle ϕ, $C(d, \phi)$. The co-occurrence matrix for an image with gray values in the range of [0, Ng-1] is defined as the following:

$$C(d,\phi) = \begin{bmatrix} P(0,\,0) & P(0,\,1) & \dots & P(0,\,N_g-1) \\ P(1,\,0) & P(1,\,1) & \dots & P(1,\,N_g-1) \\ \vdots & \vdots & \ddots & \vdots \\ P(N_g-1,\,0) & P(N_g-1,\,1) & \dots & P(N_g-1,\,N_g-1) \end{bmatrix}, \qquad (2.46)$$

where $P(g_i, g_j) = \frac{1}{R}\eta(g_i, g_j)$ is the probability for a gray value pair (g_i, g_j), $\eta(g_i, g_j)$ is

Table 2.2: Sets of features extracted from each single cell.

Feature set	**Total number**
Haralick texture	260
Zernike moments	49
Granularity	21
Object-related	8
Edge-related	3
Gray scale invariants	10
Tree-structured wavelets	2

the frequency of a gray value pair (g_i, g_j) and R is the total number of possible pixel pairs in the image depending on d and ϕ. In this work, we compute co-occurrence matrices for the distances of one to five pixels and angles of 0°, 45°, 90°, and 135°. Thirteen statistical features (*e.g.*, the angular second moment, contrast, correlation, variance, and entropy) are computed for each co-occurrence matrix, which leads to 260 image features that describe the texture of an image. Table 2.3 lists the thirteen features.

Object- and edge-related features: For the object- and edge-related features, the basic attributes of an object, such as the area (number of pixels), contour length (perimeter), and moments (*e.g.*, the mean gray value and standard deviation of the gray value), are measured. The circularity of an object is computed by $\frac{p^2}{A}$, where p is the perimeter of an object and A is the area of the object. Feret's distance, which is the longest distance within an area, is computed using the greatest possible distance between any two contour pixels. The edge-related features are computed by applying the Laplace and Sobel filters to the image and refining the detected edges with a thresholding method. The number of detected edge pixels is used as a further feature.

Granularity features: The granularity features depend on the relation of neighboring pixel pairs. The differences in the gray levels of the center pixel and all

Table 2.3: Thirteen statistical features computed on a co-occurrence matrix.

Angular second moment	$f_1 = \sum_i \sum_j (P(i,j))^2$
Contrast	$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_i \sum_j P(i,j) \right\}, \; \lvert i-j \rvert = n$
Correlation	$f_3 = \frac{\left[\sum_i \sum_j (ij) P(i,j)\right] - \mu_x \mu_y}{\sigma_x \sigma_y}$
Variance	$f_4 = \sum_i \sum_j (i-\mu)^2 P(i,j)$
Inverse difference moment	$f_5 = \sum_i \sum_j \frac{P(i,j)}{1+(i-j)^2}$
Sum average	$f_6 = \sum_{i=0}^{2N_g-2} i P_{x+y}(i)$
Sum variance	$f_7 = \sum_{i=0}^{2N_g-2} (i - f_6)^2 P_{x+y}(i)$
Sum entropy	$f_8 = -\sum_{i=0}^{2N_g-2} P_{x+y}(i) \log P_{x+y}(i)$
Entropy	$f_9 = -\sum_i \sum_j P(i,j) \log P(i,j)$
Difference variance	$f_{10} = \sum_{i=0}^{N_g-1} i^2 P_{x-y}(i)$
Difference entropy	$f_{11} = -\sum_{i=0}^{N_g-1} P_{x-y}(i) \log P_{x-y}(i)$
Information measure I	$f_{12} = \frac{H_{xy} - H_{xy}^1}{\max\{H_x, H_y\}}$
Information measure II	$f_{13} = \sqrt{1 - \exp(-2(H_{xy}^2 - H_{xy}))}$

Definition:
$H_{xy}^1 = -\sum_i \sum_j P(i,j) \log(P_x(i) P_y(j))$
$H_{xy}^2 = -\sum_i \sum_j P_x(i) P_y(j) \log(P_x(i) P_y(j))$
$P_x(i) = \sum_j P(i,j), \; P_y(i) = \sum_i P(i,j)$
$P_{x\pm y}(k) = \sum_i \sum_{j, \lvert i \pm j \rvert = k} P(i,j)$
$\mu, \; \mu_x, \mu_y; \sigma_x, \sigma_y; H_x, H_y$ are the means and standard deviations and entropies.

pixels within a given distance (*e.g.*, 1-10 pixels) in eight directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°) are computed, and the maximum difference in each direction is stored. The mean and standard deviation of the maxima over all image pixels are computed.

Gray scale invariant features: The gray scale invariant features [17] are computed by combining a pair of neighboring pixels in an image $g(x, y)$ using a simple nonlinear kernel function f that transforms the gray value into form of $f(g(x,y)) = f_1(g(x,y)) \cdot f_2(g(x+d_1, y+d_2))$, where $\mathbf{d} = [d_1, d_2]$ is a span vector for the kernel function. This function is computed for each pixel and its neighbors

in all possible directions. The summation of the resulting values yields a value being invariant to rotation. This strategy is applied to all possible positions in the image, and the results are summed for the whole image, which yields a value that is invariant to rotation of the image content. The applied kernel functions are the followings: (1) the product of the gray values ($f_1(g) = f_2(g) = g$) and (2) the product of the square roots of the gray value ($f_1(g) = f_2(g) = \sqrt{g}$). Different grey scale invariant features are computed by varying the distances (*i.e.*, the distances between the center and neighboring pixels).

Zernike moment features: These moments are commonly used to characterize distributions. In image processing, an image region is considered as a two-dimensional density function. The moment sets of different orders and with a different basis function can be used to describe the information in an image region [111]. The complex Zernike moments [165] use a set of complex polynomials that form a complete orthogonal basis that is defined over a unit circle. The image is translated and scaled to a unit disc first (disc centered at the origin (0,0) with radius one) because these Zernike polynomials defined within a unit circle. For an image $g(x, y)$, the Zernike moments can be computed using the following:

$$Z_{mn} = \frac{m+1}{\pi} \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} V^*_{mn}(x, y) g(x, y), \tag{2.47}$$

where $x^2 + y^2 \leq 1$ and $V^*_{mn}(x, y)$ is the complex conjugate of a Zernike polynomial of the degree m, n is a positive integer with $0 \leq n \leq m$ and m-n is even, and

$$V_{mn}(x, y) = \sum_{s=0}^{(m-n)/2} (-1)^s \frac{(m-s)!}{s! \left[\frac{m+n}{2} - s\right]! \left[\frac{m-n}{2} - s\right]!} \cdot (x^2 + y^2)^{\frac{m}{2}-s} \exp(yn\theta) \tag{2.48}$$

where $\theta = \tan^{-1}(y/x)$, and $y = \sqrt{-1}$. As proposed by Boland *et al.* [10], the magnitudes $|Z_{mn}|$ of the moment are used as image features. The Zernike moments are calculated up to degree 12 ($m \leq 12$) and all possible values for n; this calculation results in 49 features.

Tree-structured wavelets: A wavelet transform decomposes a signal into different frequency channels. Applying a wavelet transform to a 2D input image yields

four subimages. Each subimage comprises a part of the whole frequency bandwidth and has a quarter of the input image resolution. Daubechies wavelets [30] are widely used in signal and image processing. The decomposition is based on 12-tab Daubechies wavelets. Tree-structured wavelet transform [21] is a multiresolution analysis approach. This approach decomposes only the significant frequency channels of the subimages, that contain the most information. The information content is determined using the image energy. The image energy is computed from an energy function $E(g)$ for an image g(x,y) with x- and y-dimensions N_x, N_y, which is the mean of the absolute gray values given by the following:

$$E(g) = \frac{1}{N_x N_y} \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} |g(x,y)| \,. \tag{2.49}$$

The decomposition is recursively performed on the input image depending on the image size. At each decomposition step, the feature used is a product of the highest energy value and a constant representing the frequency channel in which the highest energy was observed.

2.2.5 Pairwise phenotypic descriptors for protein-protein interactions

The fraction and maxima features

As mentioned in Section 2.2.4, the segmentation and feature extraction were performed using an automated image processing system as described in [55]. Each single cell nuclei was segmented using the Otsu thresholding method and characterized using morphological descriptors (Table 2.2) such as the Haralick texture, Zernike moment, granularity features, object-and edge-related features, grey-scale invariants, number of cells and pixels. These features were used to distinguish between different phenotypes of the cells. Using SVM analysis (Section 2.2.3.2), each single cell was classified into the following four morphological classes: interphase, apoptosis, mitosis, and shape (cluster of cells). The classifier was trained to distinguish between the four phenotypic classes using the trained morphological classes that were manually annotated by an expert. The fractions of each phenotype were computed with respect to the number of cells in an image for each knocked-down

gene. The features of cell proliferation, median and standard deviation of the cellular intensities, were also calculated. To obtain features for a *pair* of knocked-down genes, we calculated the absolute value of the differences between the features for each gene of the respective pair. These features were termed "fraction features".

To obtain more discriminative features, we used features from the original study by Neumann and co-workers [98]. The phenotypic scores of the seven morphological phenotypes from the Mitocheck database (www.mitocheck.org) were extracted and comprised the following features: 1) mitotic delay, 2) binuclear, 3) polylobed, 4) grape, 5) large, 6) dynamic change, and 7) cell death. The scores were derived from the maximum difference of the cell counts between the negative controls and the cells of the respective class (of one of the seven morphological phenotypes). The time points for these maxima were also taken as features. We also calculated the absolute value of the differences between the features for each gene of the respective pair to obtain features for a *pair* of knocked-down genes. These features were termed "maxima features".

LDA-Similarity and proximity features

We used linear discriminant analysis (LDA) (Section 2.2.3.1) to distinguish between two sets of single cells in which each set of cells had a different gene knocked down. These features were termed *LDA-performance* features. If two sets of single cells were classified well (*i.e.*, the phenotypes that resulted from the knockdown of the corresponding two genes were dissimilar), then these sets yielded a favorable discrimination performance. The performance (accuracy) of the classification was used as a similarity feature. For the *proximity-features*, we computed the distances between a reference gene and two genes instead of computing the distance between two genes directly; we then computed the difference of these two distance vectors. If these two genes are close together, these two distances should also be close together. These *proximity features* were computed with 5 reference genes to find a vector of the distance. To obtain different feature vectors, we selected 5 reference genes that are distinct from each other. An integer linear programming problem was formulated for finding these reference genes.

The LDA-performance feature and the Euclidean distance of maxima features were used as distances. For all genes, the distance of a gene to the other genes was

computed. The problem of finding k genes that were distant from each others can be formulated as the following quadratic problem:

$$\begin{aligned} \max \quad & \sum d_{ij} x_i x_j, \\ \text{subject to:} \quad & \sum_i x_i \leq k, \\ & x_i,\, x_j \in \{0,1\}, \\ & i,j = 1,\ldots N, \end{aligned} \tag{2.50}$$

where d_{ij} is the distance from gene x_i to gene x_j. However, this problem can be transformed into an integer linear problem. We introduced a new variable y_{ij}, which can be represented as $y_{ij} = x_i x_j$. The integer linear problem is defined as the following:

$$\begin{aligned} \max \quad & \sum d_{ij} y_{ij} \\ \text{subject to:} \quad & \text{(I)} \quad y_{ij} \leq x_i, \\ & \text{(II)} \quad y_{ij} \leq x_j, \\ & \text{(III)} \quad y_{ij} \geq x_i + x_j - 1, \\ & \text{(IV)} \quad \sum_i x_i \leq k, \\ & x_i \in \{0,1\},\, y_{ij} \in \{0,1\}, \\ & \forall i,j \in \{1,\ldots,N\},\, i < j. \end{aligned} \tag{2.51}$$

In this maximization case, the constraint (III) can be discarded. To reduce the complexity of the problem, thirty genes with distances that deviated significantly were selected. In this study, we performed computations to obtain five optimal genes (k=5). To find the difference between the distances of optimal gene pairs and other gene pairs, the distances were normalized for values in a range from 0 to 1, and the distances of all combination of each sets were computed. When computations were performed using the LDA-performance and maxima features as the distances, the distances of all combinations of optimal gene pairs yielded significantly higher values than the distance of all combinations of other gene pairs (Figure 2.9).

In summary, the pairwise phenotypic features are comprised of the following sets of fraction features, LDA-performance feature, maxima features, and proximity features. The complete list of pairwise phenotypic features is shown in Table 2.4.

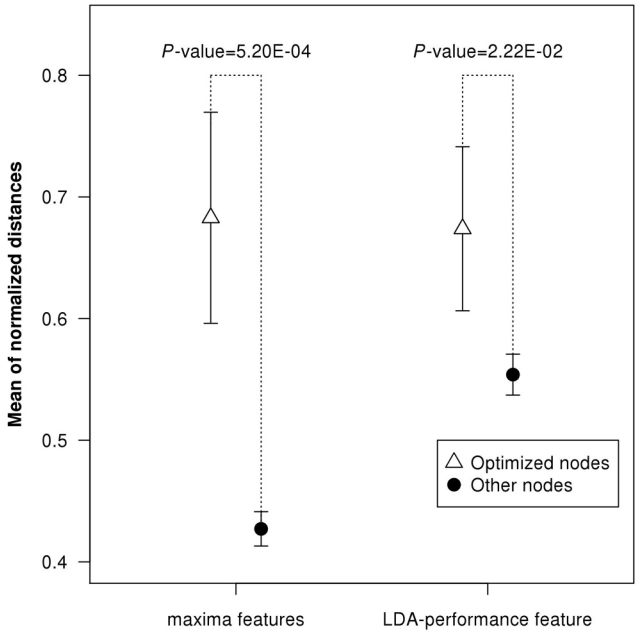

Figure 2.9: Comparison of the distances between all combinations of the optimal genes and all combinations of the other genes. The *P*-values were calculated for the two distributions of distances between the five optimal genes and the other genes using a Wilcoxon test.

Table 2.4: Pairwise phenotypic descriptors

Feature name	**Description**
Fraction features	
frApop	Distance computed from the fractions of Apoptotic cells
frInter	Distance computed from the fractions of Interphase cells
frMito	Distance computed from the fractions of Mitosis cells
frShape	Distance computed from the fractions of Cluster cells
medMean	Distance computed from the medians of mean of cell intensities
medSD	Distance computed from the medians of standard deviation of cell intensities
medNbPixel	Distance computed from the medians of number of cell images pixels
medNumCell	Distance computed from the medians of number of cells
ProliferRate	Distance computed from the cell proliferation rates

Continued on next page...

Table 2.4 – continued from previous page

Feature name	Description
LDA-performance feature	
LDA-performance	LDA similarity
Maxima features	
MitoticDelay	Distance computed from the maximum scores of a Mitotic Delay phenotype
Binuclear	Distance computed from the maximum scores of a Binuclear phenotype
Polylobed	Distance computed from the maximum scores of a Polylobed phenotype
Grape	Distance computed from the maximum scores of a Grape phenotype
Large	Distance computed from the maximum scores of a Large phenotype
DynamicChange	Distance computed from the maximum scores of a Dynamic Change phenotype
CellDeath	Distance computed from the maximum scores of a Cell death phenotype
tMitoticDelay	Distance computed from the time point with the max.score of a Mitotic Delay phenotype
tBinuclear	Distance computed from the time point with the max.score of a Binuclear phenotype
tPolylobed	Distance computed from the time point with the max.score of a Polylobed phenotype
tGrape	Distance computed from the time point with the max.score of a Grape phenotype
tLarge	Distance computed from the time point with the max.score of a Large phenotype
tDynamicChange	Distance computed from the time point with the max.score of a Dynamic Change phenotype
tCellDeath	Distance computed from the time point with the max.score of a Cell death phenotype
Proximity features	
SpScoreRef(1-5)	Distance computed from reference gene(1-5) using the maxima features
LDAperfRef(1-5)	Distance computed from reference gene(1-5) using the LDA-performance

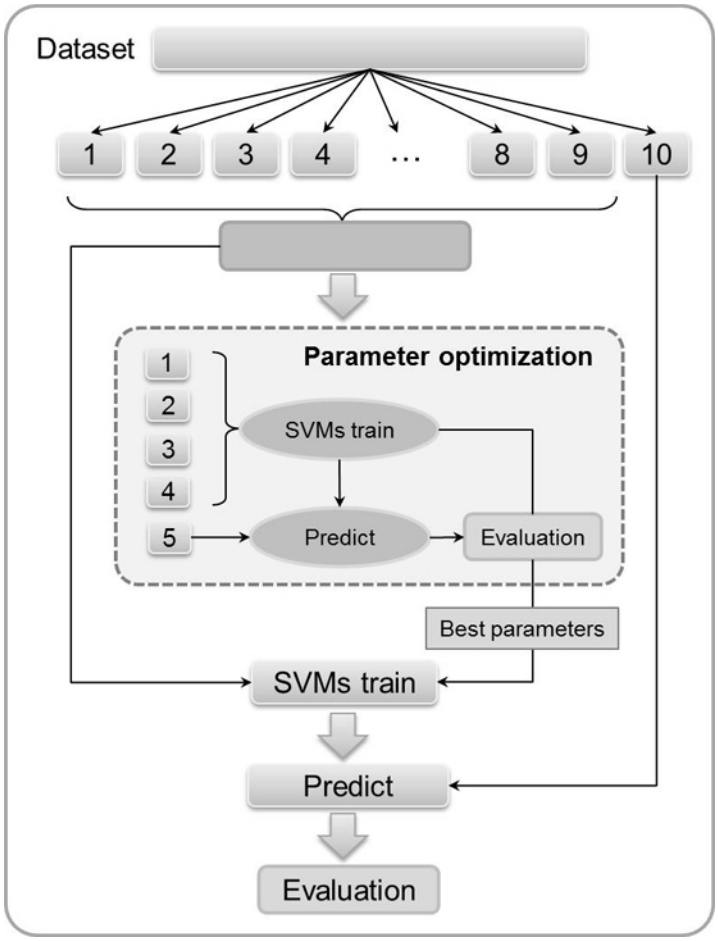

Figure 2.10: Two nested cross-validation loop. The 10-times 10-fold cross-validation technique was performed with a 5-times 5-fold cross-validation for parameter optimization.

2.2.6 Classification of interactions with a role in the activation or inhibition of signal transduction

Based on the pairwise phenotypic descriptors, we classified the interactions as having a role in the activation or inhibition of signal transduction using the SVM as mentioned in Section 2.2.3.2. To assess the performance of the classifiers, 10-times-10-fold cross-validations were performed (Figure 2.10). In each cross-validation, the PPIs involved in the activation and inhibition of signal transduction were randomly split into ten equally sized, non-overlapping subsets. The nine subsets were concate-

nated and used for training the classifiers and testing of the one remaining subset. The performance was measured on the test set by comparing the predictions with the true class labels. To measure the performance of the predictions of the PPIs involved in the activation of signal transduction, these PPIs were set as the positives. This process was repeated ten times until each subset was tested once. In our dataset, the sizes of the two classes (Act-PPI and Inh-PPI) differed considerably. Therefore, a data stratification was performed using an ensemble machine learning technique. In each training subset, ten SVM classifiers were trained with equal stratified numbers of randomly selected PPIs that are involved in the activation and inhibition of signal transduction.

To optimize parameters for the classifiers, the nested cross-validation loops were employed. In the inner loop, the cross-validations were repeated to obtain the optimal parameter set (Figure 2.10). It is crucial that the test data were not included in this inner cross-validation. For each combination of parameters used for the training step, the cross-validation performance was measured, and the significant parameters were selected in this inner loop. In this work, we used a radial basis function as a kernel for our SVM. Therefore, there are two parameters for the optimization of the kernel (Section 2.2.7). To obtain the overall performance of the classifiers from the nested cross-validation loops, we repeated the cross-validation procedure 10 times. The votes of each testing sample were summed from the predictions of the classifiers for a certain class. Using these votes, a receiver operator characteristics curve (ROC curve) was used to measure the performance of the classifiers (Section 2.2.8). The performance was estimated by the area under the curve (AUC) for the entire range of thresholds based on the votes. To predict new interactions, all 1000 trained classifiers were employed as an ensemble classifier that used a voting scheme in which each SVM contributed one vote. Figure 2.10 shows the cross-validation procedure with a 10-fold cross-validation for the outer loop and a 5-fold cross-validation for the inner loop. The software library LIBSVM [20] was used for the SVM classifications.

2.2.7 Parameter optimization and voting scheme technique

The SVM algorithm (Section 2.2.3.2) was employed for the classification. Using a radial basis kernel, there are two SVM parameters that can be optimized: 1) the regularization term that defined the costs of false classification (C); and 2) kernel width parameter (γ), which regulates the variance of the kernel. Using an approach

proposed by Hsu *et al.* [59], we performed a grid search on the training data to maximize the classification accuracy on a defined parameter space. The parameter space was defined using the value of C and γ, which grow exponentially ($C = 2^n$, $\gamma = 2^n$, n = -5, -4,..., 4, 5). To measure the classification performance, we split the training data into two parts. One part is used for the training data (also called the training set) that is used to train the classifier. The other part is called a test set and is used for the testing of the trained classifier. The classification performance is measured on the test set by comparing the classifier output with the true classes of the test set. Then, the percentage of correct classifications can be determined. However, training and testing the data on a training set might not reflect the true classification performance and produce poor classification results for other data because of the use of a specific training set. To achieve more reliable results, the parameter testing can be performed on several independent data sets using the cross-validation technique. In this work, we used a 5-fold cross-validation for the parameter optimization (Figure 2.10). The cross-validation technique splits the data into 5 subsets of equal size. Four of the subsets were used for training the classifier and the other subset is used to test the classifier. This process is repeated 5 times until each subset has been tested once. The best determined combination of the C and γ can be used for the whole training set to train the final classifier.

2.2.8 Performance measurements

By comparing the predictions with the true classes, we can generate a confusion matrix, which is also called contingency table. Table 2.5 shows an example of a confusion matrix of a two-class classification task. TP are the true positives, FN are the false negatives, FP are the false positives, and TN are the true negatives. *Accuracy* is a commonly used classification measurement. The accuracy measures the proportion of correct predictions:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN}. \tag{2.52}$$

From the confusion matrix, we can compute other performance values such as the sensitivity, specificity, positive predictive value and negative predictive value. The sensitivity or recall is the proportion of actual positives that are correctly classified, whereas the specificity is the proportion of negatives that are correctly classified.

Table 2.5: Confusion matrix of the two-class classification

		Predicted Classes	
		Positive	Negative
True Classes	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

The positive predictive value or precision rate measures the proportion of correct positive predictions performed by a classifier, whereas the negative predictive prediction denotes the portion of correct negative predictions:

$$\text{sensitivity} = \frac{TP}{TP + FN}, \tag{2.53}$$

$$\text{specificity} = \frac{TN}{TN + FP}, \tag{2.54}$$

$$\text{positive predictive value} = \frac{TP}{TP + FP}, \tag{2.55}$$

$$\text{negative predictive value} = \frac{TN}{TN + FN}. \tag{2.56}$$

These performance measures resulted from a dataset that is called a test set. The test set with known class labels is the data remaining after the data of the training set is removed. We perform the measurement on the test set instead of the training set to avoid the overestimation the measurement. The test set is applied to the trained classifier and predicts the class labels. The predicted labels are then compared with the true labels and the performance measurements are calculated as described above.

Receiver operator characteristics and the area under the ROC curve

A common approach used to compute the overall classification performance is the receiver operator characteristic (ROC) curve and the area under the ROC curve (AUC). The ROC is suitable for measuring the performance of a classifier system using various thresholds of stringency (*e.g.*, when using voting scheme technique). The ROC curve shows the true positive rate (sensitivity) versus the false positive rate (1-specificity). The overall performance of the classifiers is calculated from the area under the ROC curve. A perfect classifier has an AUC of 1.0, whereas random guessing produces an AUC of 0.5. Figure 2.11 ROCs shows an example of an ROC

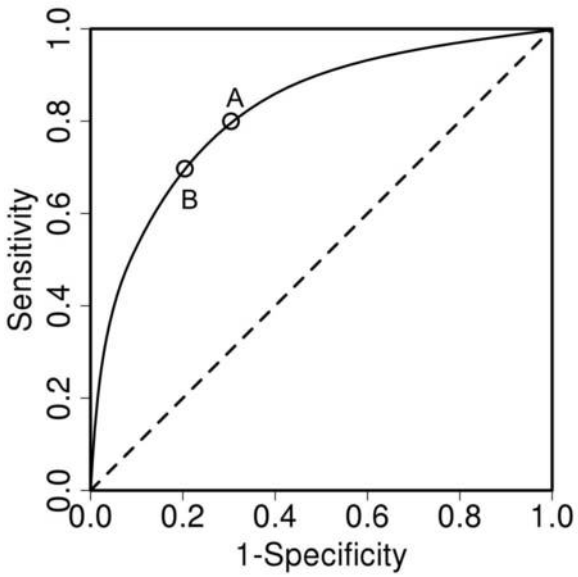

Figure 2.11: An example of an ROC curve. The performance of multiple thresholds can easily be investigated by plotting. For example, a certain threshold at point A yields a sensitivity of 0.8 and a specificity of 0.7. Another threshold at point B yields a sensitivity of 0.7 and a specificity of 0.8.

curve.

By changing the thresholds, we determined a point on the ROC curve. From the example curve of Figure 2.11, we found the point A using a certain threshold. At this point, we found a sensitivity value of 0.8 and a specificity value of 0.7. Using a different threshold, we found the point B, which has a sensitivity value of 0.7 and a higher specificity value of 0.8. The dashed diagonal line represents the results from random predictions that produce an AUC of 0.5. The perfect classifier yields a curve that includes the coordinate (0,1) in the upper left corner, which corresponds to 100% sensitivity, 100% specificity and an AUC that is equal to one.

2.2.9 Consistency score

To improve the precision of classification performance, we conducted a statistical post-processing step that was used to filter our results. We compared the effect of each of the down-regulated genes of a pair (gene_i, gene_j) to all other investigated genes (gene_k). If both knocked-down genes (*i.e.*, the gene i and j) showed the same

prediction to the other gene k, we defined the knocked-down genes as interacting "consistently" with respect to gene_k. Similarly, if one of them showed a prediction of activation and the other inhibition, we set that pair to be inconsistent with respect to gene_k. This was performed for all other genes k, $k \in \{all\,genes \setminus \{i, j\}\}$, and the portion of consistent and inconsistent interactions was used to define the consistency score (high consistency = a higher number of other genes that show the same activation/inhibition predictions to both genes of the pair). This criterion was used to filter out gene pairs that had a high consistency but a low number of votes used for the prediction of a PPI as involved in the activation of signal transduction (and *vice versa* for inhibition). If the voting score of a gene pair was less than 100, it was predicted to be a PPI that is involved in the inhibition of signal transduction; In turn, if the vote of a pair was more than 900, it was defined as a potential PPI that is involved in the activation of signal transduction. All other predictions were assigned as undefined. We computed the consistency score from the percentage of consistency and inconsistency values. To quantify the difference between the consistency and inconsistency values, we calculated the similarity score by transforming the percentage of consistency and inconsistency values into the range between -1 and 1 using the following hyperbolic tangent function:

$$f(x) \;=\; \tanh(k * X), \tag{2.57}$$

where X is the proportion of consistency values subtracted by the proportion of inconsistency values. For this study, we used the optimized parameter k=5 and we improved the negative (or lower than average) and positive (or higher than average) consistency scores to yield the PPIs as involved in the inhibition and activation, respectively, of signal transduction.

2.2.10 Enrichment tests for the consistency score of protein pairs

All protein pairs and their consistency score (gp_i, i=1,..., all protein pairs) were investigated using the interaction databases, STRING version 9.0 [136] and MetaCore$^{\text{TM}}$ version 6.8, www.genego.com. We applied the gene set enrichment analysis (GSEA) strategy of Subramanian *et al.* [131]. The goal of GSEA is to determine whether the evidenced interactions (a list S with N_H pairs in the database) are randomly distributed throughout our ranked consistency scores $r(gp_j) = r_j$ in a

list $L = \{gp_1, gp_2, ..., gp_N\}$ or found at the top or bottom of the list. This approach is essentially a Kolmogorov-Smirnov test of running sums over the ranked scores. The enrichment score (ES) was computed and indicated the degree of overrepresentation of a set S in the top or bottom of the ranked list L. The algorithm walks into the ranked list L, and a running-sum is increased if the gene pairs found in the list S, are also found in the database; otherwise, the running-sum is decreased. The maximum deviation between the zero encountered in the random walk and the magnitude of the increment was calculated as the ES. The ES is calculated as follows:

$$ES(S) = \max_{1 \leq i \leq N} \left\{ \sum_{\substack{gp_j \in S \\ j \leq i}} \frac{1}{N_R} - \sum_{\substack{gp_j \notin S \\ j \leq i}} \frac{1}{N - N_R} \right\} \tag{2.58}$$

where N_R is the number of $gp_j \in S$. The ES is the fraction of interaction pairs in S running up to i, and the value is penalized by the fraction of the interaction pairs not in S running up to i. To assess the significance of the ES, we compared the observed ES with the null set of ES scores that were computed using a permutation-based approach. We found the null distribution of the permutated ES by permuting the interaction pair labels and re-computing the $ES(S)$. We repeated this step 10^4 permutation times and computed the nominal P-value for S from the null distribution. The nominal P-value is estimated as the portion of the permutated ES which is greater than the observed ES.

Chapter 3

Results

This chapter describes our results mainly consisting of two parts. Section 3.1 describes the results of clustering of cells infected with Hepatitis C Virus to identify host factors. Section 3.2 presents the results of characterizing the activities of protein-protein interactions using machine learning approaches.

3.1 Clustering of cells infected with Hepatitis C Virus

Viruses can spread within a host through the release of cell-free virions or direct passage between infected and non-infected cells. In general, direct cell-cell transfer is considerably more efficient than cell-free transfer [141] and can be supported by filopodial bridges [124], virological synapses, or nanotubes [120]. As a consequence of such a viral cell–cell spreading, clusters of infected cells may be formed. It was recently reported that the spatial distribution of cells can influence infection behavior. Snijder and co-workers observed intriguing relationships between virus species, the spatial distribution of cells and infection rates. While the infection efficiency of a rotavirus was considerably increased in sparse populations, Dengue Viruses mainly employed cells located at the edges of islets, and murine hepatitis viruses were preferably found in dense cell populations [126]. To analyze such clustering patterns systematically, statistical methods for point pattern analysis can be employed. Section 3.1.1 reports the parameter optimization of the clustering method

and shows our hits compared with primary and secondary experimental screens. Section 3.1.2 provides a functional interpretation of our hits. Section 3.1.3 provides a comparison of the clustering behavior of cells infected with HCV and cells infected with Dengue Virus.

3.1.1 Parameter optimization, choice of the most suitable clustering analysis method and assembly of significant hits

We identified cellular protein kinases involved in HCV replication by observing the replication and clustering of infected cells upon silencing of protein kinases (2157 siRNAs targeted 719 human protein kinase genes). Virus-infected cells were identified through viral GFP expression observed using fluorescence microscopy analysis. Host siRNA hits were identified based on three different approaches, (i) using the viral GFP fluorescence intensity of the primary screen, (ii) the luciferase intensity of the secondary screen and (iii) the clustering analysis method. In applying the clustering analysis method, we computed a z-transformed clustering score for all knockdowns. We analyzed the clustering of infected cells using the DAPI channel (nucleus staining) to define the center of mass and the viral GFP signal for label-

Table 3.1: Pearson's correlation coefficients for the intensity values of the scores from K-function and the standard readouts (intensity values of the primary and secondary screens).

	K-function (Inhomogeneous)				**K-function (Homogeneous)**
	40%	35%	30%	25%	35%
Correlation with intensities of the primary screen	0.51	0.55	0.49	0.36	0.36
Correlation with intensities of the secondary screen	0.31	0.32	0.34	0.28	0.23

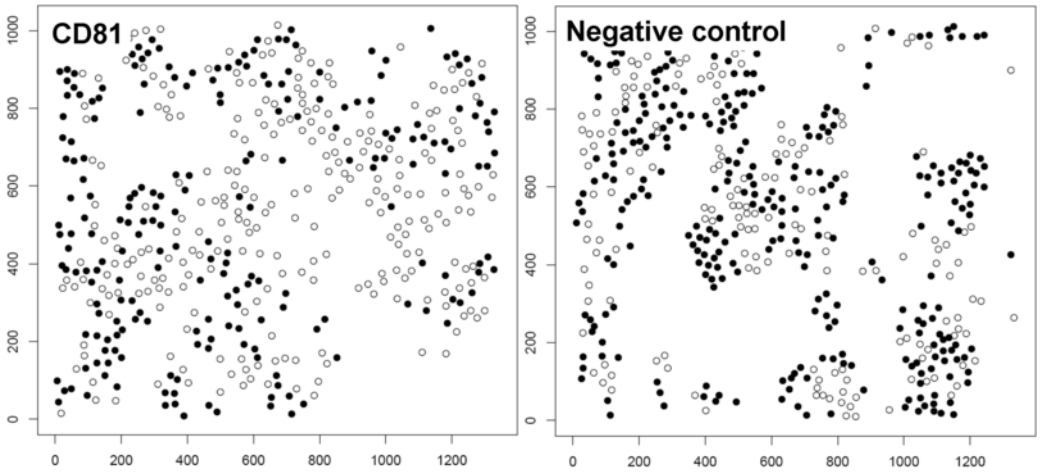

Figure 3.1: Images of positive (knockdown of CD81) and negative controls (non-silencing siRNA). Knockdown of CD81 resulted in a rather random distribution of infected cells (black dots), while infected cells were highly clustered when no genes were silenced (unhindered viral replication).

ing the cells as infected and non-infected. Low clustering scores were yielded if the infected cells did not cluster, while high values resulted specifically if the infected cells showed high clustering. This trend is demonstrated exemplarily in Figure 3.1.

To detect the clustering we used K-function and optimized the performance by varying the range of the radius. As the objective function, we analyzed the correlation of the z-scores from K-function for all knocked-down genes with the z-scores from the intensity readout of the primary screen and secondary screen. Table 3.1 shows the obtained results. The best correlation with the primary screen was 0.55 using a radius range of 35%.We investigated the performance of a well-established clustering analysis method, Quadrat Analysis [156]. Quadrat Analysis was tested, and the correlation with the primary and secondary experimental screens was observed. We optimized the parameters for the Quadrat Analysis (QA, Section 2.1.4) by varying the number of rows (i) and columns (j), with $i = 3$, 4, 5 and $j = 4$, 5, 6, yielding different grid sizes. Pearson's correlation coefficients were computed from VMR scores and intensity readout values from the primary and secondary experimental screens. The results are presented in Table 3.2. However, the method

Table 3.2: Pearson's correlation coefficients for the intensity values of the Quadrat Analysis and the standard readouts (intensity values of the primary and secondary screens)

	Quadrat Analysis		
	QA4x3	**QA4x4**	**QA4x5**
Correlation with intensities of the primary screen	0.25	0.23	0.23
Correlation with intensities of the secondary screen	-0.02	0.029	-0.027
	QA5x3	**QA5x4**	**QA5x5**
Correlation with intensities of the primary screen	0.23	0.23	0.22
Correlation with intensities of the secondary screen	0.027	-0.0018	-0.033
	QA6x3	**QA6x4**	**QA6x5**
Correlation with intensities of the primary screen	0.22	0.20	0.19
Correlation with intensities of the secondary screen	-0.07	-0.04	-0.097

showed less correlation with the intensity readouts (Table 3.2 shows the results for several parameter settings).

Additionally, the homogeneous K-function was inferior to the inhomogenous K-function (the result with the best radius range is given in Table 3.1). Here, we report results using the inhomogenous K-function with the optimized parameter (radius range = 35%). Knockdown of CD81 gene (positive control) resulted in low clustering of the infected cells, while the negative control (non-silencing siRNAs) showed a comparably high tendency for infected cells (black dots) to cluster. The clustering scores were -2.3 and 2.2 for CD81 and the negative control, respectively. In the primary screen, the mean intensities of viral GFP were calculated for each

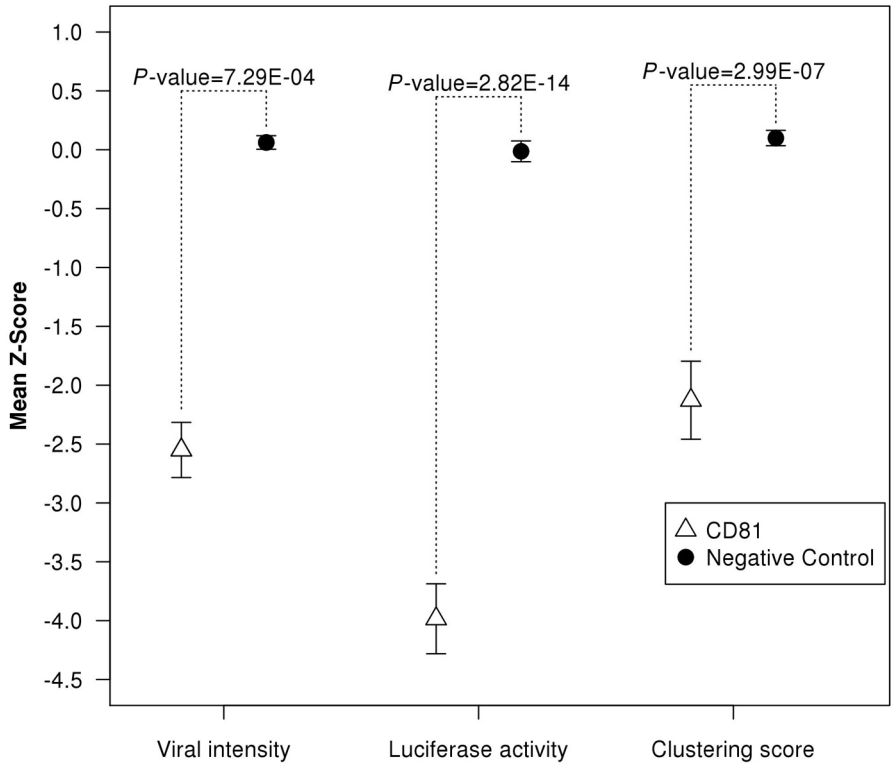

Figure 3.2: Comparison of the scores of the positive control (CD81) and the negative controls (non-silencing siRNA) for all applied methods. The P-values were calculated for the two distributions of CD81 and the negative controls using Student's t-test.

knockdown and replicate (12 replicates), following which their z-scores were computed with respect to the bulk of the data, and genes with significant low z-scores were selected (P-value < 0.05). Significant genes were defined similarly based on the secondary screen. The difference between the z-score distributions of the positive control (CD81) and the negative controls is shown in Figure 3.2. The separation of the distributions shows CD81 to be a significant down regulator in all three approaches (i.e., primary screen, secondary screen and clustering analysis method). The numbers of significant hits and their intersections are summarized in Figure 3.3. Observation of viral signal intensities in the primary screen yielded 85 significant genes. A total of 178 genes selected from the primary screen were observed with the secondary screen yielding 64 significant genes. The clustering analysis method

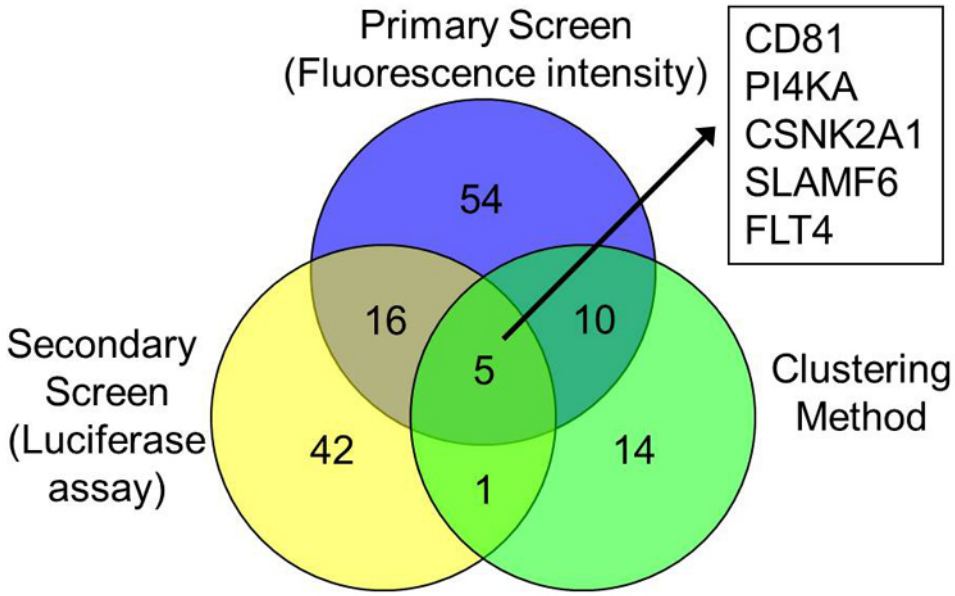

Figure 3.3: Venn diagram of the hits for all three applied methods.

yielded 30 genes (shown in Table 3.4). All three positive controls showed significantly low clustering scores (CD81: P-value = 6.61E-07; HCV-321: P-value = 1.53E-13; HCV-138: P-value = 1.20E-10). Five genes were found to be significant with all three methods: CD81, PI4KA, CSNK2A1, SLAMF6 and FLT-4 (Table 3.3). Note that the positive controls HCV-321 and HCV-138 were not used in the secondary screen. CD81 was used as a positive control, as it is well-known viral receptor of HCV [166] and is involved in HCV entry [114].

3.1.2 Functional interpretation of the results

In addition to CD81, we detected four host factors as significant using all three analysis approaches (PI4KA, CSNK2A1, SLAMF-6 and FLT-4). Phosphatidylinositol 4-kinase-α (PI4KA) is well known to be required for HCV replication [5, 11, 77, 137, 142, 145]. It was shown *in vitro* that Casein kinase II (for which CSNK2A1 encodes the subunit alpha) phosphorylates the non-structural HCV protein NS5A [70]. Fms-related tyrosine kinase 4 (FLT-4), also known as vascular endothelial growth factor receptor 3 (VEGFR-3), is a member of the tyrosine kinase

Table 3.3: Host factors detected with all three analysis methods

Entrez gene ID	**Gene symbol**	**Gene name**	***P*-value (Clustering method)**
975	CD81	CD81 molecule	6.61E-07
5297	PI4KA	Phosphatidylinositol 4-kinase, catalytic, alpha	0.0019
1457	CSNK2A1	Casein kinase 2, alpha 1 polypeptide	0.0274
114836	SLAMF6	SLAM family member 6	0.0345
2324	FLT4	Fms-related tyrosine kinase-4	0.0445

receptor family. Over-expression of the short splice variant of VEGFR-3 stimulated cell growth in HepG2 cells [79], which may favor infectious spreading of the virus. Interestingly, a retrovirus was found to be integrated into an intron of FLT-4 in the genome which may have resulted in an evolutionary advantage for this virus [61]. SLAMF-6 belongs to the signaling lymphocytic activation molecule family and is a transmembrane receptor mainly expressed in natural killer (NKT) cells. This receptor serves as a docking site for several signaling molecules [38, 147]. It was shown that SLAMF-1 and SLAMF-6 critically control the characteristic expansion and differentiation of NKT cells following thymic selection [49]. SLAMF-6 may be a suitable interesting candidate for investigating the uptake and signal propagation of the virus during its entry into the host cell.

3.1.3 Comparing the clustering behavior of HCV and the Dengue Virus infection

The same experimental set-up as was used for HCV was applied to observe cells infected with the Dengue Virus (DV) [92]. It is known that DV infects the edges of islets of cell populations rather than forming clusters of infections [126]. We also observed this behavior in our data which is shown exemplarily in Figure 3.4. We compared the clustering scores for non-silencing siRNA images for both datasets and

Table 3.4: The first 30 candidate genes from the clustering analysis approach.

Entrez Gene ID	**Gene Symbol**	**Gene Name**	***P*-Value**	**Mean *Z*-score**
	HCV_321	Positive Control	1.53E-13	-2.9217
	HCV_138	Positive Control	1.20E-10	-2.6395
975	CD81	CD81 molecule (Positive Control)	6.61E-07	-2.1275
7852	CXCR4	Chemokine (C-X-C motif) receptor 4	6.14E-05	-0.6285
5297	PI4KA	Phosphatidylinositol 4-kinase, catalytic, alpha	0.0019	-3.5111
4233	MET	Met proto-oncogene (hepatocyte growth factor receptor)	0.0030	-0.8307
10298	PAK4	p21 protein (Cdc42/Rac)-activated kinase 4	0.0041	-0.5306
51447	IP6K2	Inositol hexakisphosphate kinase 2	0.0061	-0.6380
10188	TNK2	Tyrosine kinase, non-receptor, 2	0.0072	-0.7748
9212	AURKB	Aurora kinase B	0.0131	-0.5962
2645	GCK	Glucokinase (hexokinase 4)	0.0175	-0.6753
5586	PKN2	Protein kinase N2	0.0193	-0.6841
140767	NRSN1	Neurensin 1	0.0197	-0.3436
440275	EIF2AK4	Eukaryotic translation initiation factor 2 alpha kinase 4	0.0203	-0.7116
659	BMPR2	Bone morphogenetic protein receptor, type II (serine/threonine kinase)	0.0238	-0.6405
5298	PI4KB	Phosphatidylinositol 4-kinase, catalytic, beta	0.0259	-0.6006
6198	RPS6KB1	Ribosomal protein S6 kinase, 70kDa, polypeptide 1	0.0274	-0.5495
1457	CSNK2A1	Casein kinase 2, alpha 1 polypeptide	0.0274	-0.6259
255239	ANKK1	Ankyrin repeat and kinase domain containing 1	0.0323	-0.6443
5605	MAP2K2	Mitogen-activated protein kinase kinase 2	0.0328	-0.6364
132158	GLYCTK	Glycerate kinase	0.0340	-0.5697
114836	SLAMF6	SLAM family member 6	0.0345	-0.4253
30849	PIK3R4	Phosphoinositide-3-kinase, regulatory subunit 4	0.0367	-0.9770
80216	ALPK1	Alpha-kinase 1	0.0408	-0.8162
51678	MPP6	Membrane protein, palmitoylated 6 (MAGUK p55 subfamily member 6)	0.0413	-0.5960
130497	OSR1	Odd-skipped related 1 (Drosophila)	0.0434	-0.5213
2324	FLT4	fms-related tyrosine kinase 4	0.0445	-0.5479
10256	CNKSR1	Connector enhancer of kinase suppressor of Ras 1	0.0448	-0.3241
5584	PRKCI	Protein kinase C, iota	0.0449	-0.3953
548596	CKMT1A	Creatine kinase, mitochondrial 1A	0.0463	-0.6092

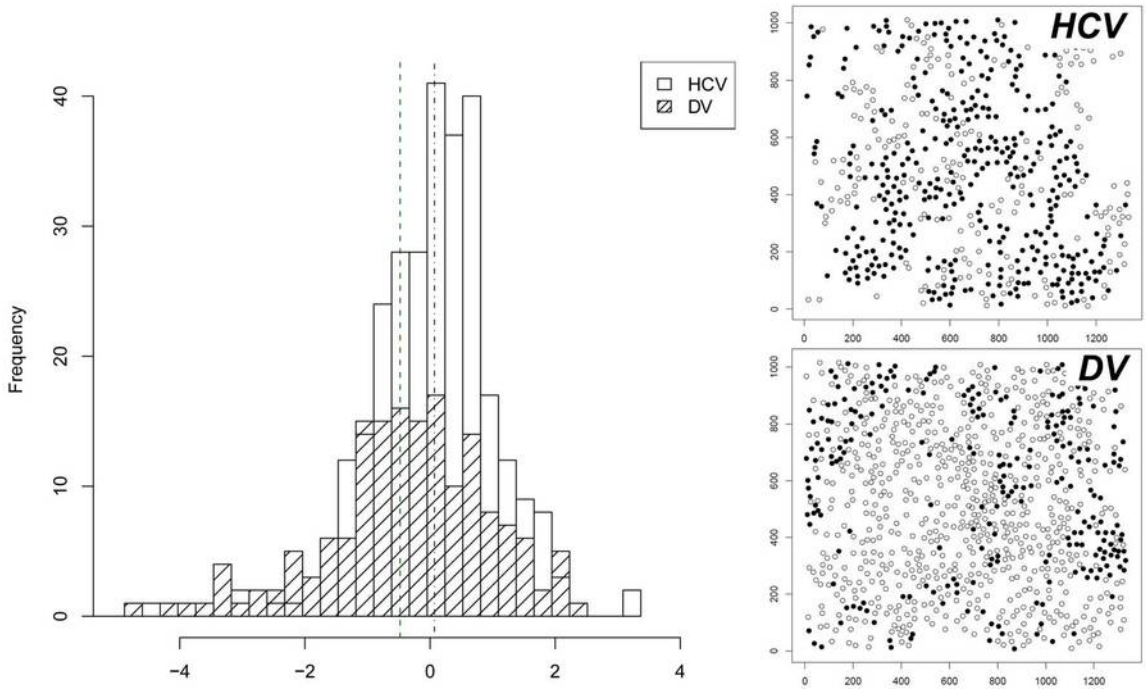

Figure 3.4: **Left**: Distribution of clustering scores for Hepatitis C Virus (HCV) and Dengue Virus (DV). In comparison with DV, cells infected by HCV showed significantly higher clustering scores (Wilcoxon test P=4.8E-04). **Right**: Typical examples of real position images of infected (filled) and non-infected cells (not filled). HCV infected cells show cluster formation, while DV infected cells populated rather the edges of cell colonies.

observed significantly higher clustering scores for cells infected with HCV (Wilcoxon test P= 4.8E-04, see Figure 3.4 for the distribution of all scores for both data sets).

3.2 Characterization of the signaling interactions

We investigated eleven signaling pathways which had a high overlap with cytokine receptors, such as the endocrine signaling, cell growth and death and the immune system pathways (Table 2.1). A total of 663 proteins for which we had phenotypic data were investigated. Among these proteins, we obtained 1927 and 676 known

activation and inhibition interactions, respectively. Gene pairs for all sets (activating protein-protein interactions (Act-PPI), inhibiting protein-protein interactions (Inh-PPI) and putative interactions with no information regarding activation or inhibition (Undef-PPI)) were analyzed. We calculated a first set of features by applying our new concept for feature generation using Linear Discriminant Analyses (LDAs). For each gene pair, gene_i and gene_j, the task of the classifier (LDA) was to distinguish images of cells with knockdown of gene_i from images of cells with a knockdown of gene_j. The performance of the LDAs served as a similarity criterion. Features describing the performance of the LDAs ((1) *LDA-performance features*) were calculated. Good performance resulted in *e.g.*, high accuracy indicating that the phenotypes of the two knockdowns were dissimilar hinting at an inhibiting interaction. In contrast, weak performance indicated similar phenotypes (hinting at an activating interaction). As additional features, we employed (2) *phenotype-fraction* features derived from counting cells according to the distinct phenotype classes of interphase, mitosis and apoptosis, and the overall cellular proliferation rates; (3) *maxima features*, *i.e.*, the time-point and height of phenotype maxima (maxima features were obtained from the original Mitocheck study, [98]); and (4) *proximity-features* calculated from the distances to well-defined reference genes within a PPI interaction network. These features are described in Section 2.2.5 and were used to learn a second set of classifiers (Support Vector Machines, SVMs) to classify the Act-PPI and Inh-PPI sets. Gene pairs from the Act-PPI and Inh-PPI sets were trained for the classifiers and their performance was assessed in an independent validation set. The trained machines (trained on the training sets) were used to define a similarity measure (consistency score). This score was high for a pair of proteins if their interactions with other proteins were of a similar nature (similar profile, with both proteins predicted to exhibit either an activating or inhibiting interaction with the other proteins) and low otherwise (showing a rather different activation/inhibition profile). Using this score, the performance was improved. The readily trained classifiers were subsequently used to predict the nature of interactions from the set of non-defined PPIs. Finally, we applied the consistency scores in a detailed analysis of cytokine receptor signaling.

3.2.1 Assembly of known activating, inhibiting and non-defined interactions

Three (non-overlapping) sets of interactions were defined. Set no. 1 consisted of well-known interactions that were described as activating. They were taken from a literature-based data repository (the Kyoto Encyclopedia of Genes and Genomes, KEGG [66, 67, 155]) and used as a reference, or gold standard for activating PPIs (Act-PPI). Set no. 2 was taken as inhibiting PPIs (Inh-PPI) and comprised pairs of genes encoding proteins for which an inhibiting interaction has been reported (listed in KEGG). Set no. 3 consisted of putative interactions for which there is no information regarding activation or inhibition (Undef-PPI, undefined PPI). This list was assembled from computationally inferred high potential interactions and from entries in a well-curated database (MetaCore, unspecific interaction). To further restrict to proteins pairs that were very likely to interact, we used these potential interacting pairs only if their protein domains were predicted to interact (protein-domain interactions were obtained from a database [161]). With this, we selected 727 non-defined interactions, which served as a basis for new predictions of the nature of their interactions (activation/inhibition).

3.2.2 Quantifying cell phenotypes

Quantitative profiles of knockdown gene images were generated using an automated image processing system [55]. Each cell nucleus was segmented using Otsu thresholding and characterized based on morphological descriptors, *e.g.*, Haralick texture, Zernike moment, granularity features, object-and edge-related features, grey-scale invariants, and numbers of cells and pixels (Section 2.2.4). These features were used to distinguish different phenotypes of cells. Each single cell was classified into one of four morphological classes: interphase, apoptosis, mitosis, or shape (cluster of cells) using a Support Vector Machine (SVM). The classifier learned to distinguish the four phenotypic classes from trained morphological classes that were manually annotated by an expert. Therefore, the fractions of each phenotype based on the number of cells were computed for each knocked-down genes. Ordinary features, *e.g.*, cell proliferation and median and standard deviation of the cell intensities, were also calculated. This feature is termed a "fraction feature". In addition, we also

generated other features, called LDA-performance, maxima, and proximity features (Section 2.2.5), for our pairwise phenotypic features.

3.2.2.1 Cell phenotype classification

We assigned four classes of cellular phenotypes: (1) interphase, (2) mitosis, (3) apoptosis (cell death phenotypes), and (4) shape (clustered nuclei). The total number of manually classified cell objects was 775. The number of cells per class is provided in Table 3.5. The cell nuclei were characterized automatically from multicellular images using the segmentation approach, and image features were extracted.

Table 3.5: Number of single cell images separated for training and testing.

Classes	**Training set**	**Test set**	**Total**
Interphase	252	62	314
Mitosis	172	43	215
Apoptosis	89	22	111
Shape	108	27	135
Total	621	154	775

We split the available samples for each class randomly in training data and testing data at a ratio 4:1, resulting in a training set size of 621 and a test set size of 154. We trained an SVM classifier with a Gaussian radial basis function (RBF) kernel on the training dataset. The samples of the test set were classified into the four classes, *i.e.*, interphase, mitosis, apoptosis, and shape. We repeated the classification step applying ten times random sampling on the whole dataset. The performances of the classification for the training and test sets are shown in Table 3.6 and Table 3.7, respectively. We yielded an overall accuracy of the training set of 99.62% and of the test set of 96.62%. Misclassifications mainly occurred between the classes mitosis and apoptosis, which are difficult to seperate, even through human identification.

Table 3.6: Confusion matrix for SVM classification of training sets based on Table 3.5. The overall accuracy is 99.62% (618.7/621).

True class	Classifier output				Accuracy
	Interphase	Mitosis	Apoptosis	Shape	
Interphase	**252**	0	0	0	**100.00%**
Mitosis	0	**172**	0	0	**100.00%**
Apoptosis	0	2.1	**86.9**	0	**97.64%**
Shape	0.2	0	0	**107.8**	**99.81%**

Table 3.7: Confusion matrix for SVM classification of test sets based on Table 3.5. The overall accuracy is 96.62% (148.8/154).

True class	Classifier output				Accuracy
	Interphase	Mitosis	Apoptosis	Shape	
Interphase	**61.1**	0	0.9	0	**98.55%**
Mitosis	0	**41.9**	1.1	0	**97.44%**
Apoptosis	0	3	**18.9**	0.1	**85.91%**
Shape	0	0	0.1	**26.9**	**99.63%**

3.2.2.2 Characterization of the phenotypic similarity and dissimilarity of cells using LDAs

For each of the 663 genes from the selected cytokine signaling pathways, cell images of each pair of genes were compared. The approach of using LDAs to describe phenotypic similarity is exemplarily described for three sample knockdowns illustrated in Figure 3.5. Two of the genes, frizzled family receptor (*FZD7*) and dishevelled 2 (*DVL2*), are closely functionally related. *DVL2* is activated by *FZD7* in the Wnt signaling cascade [160]. Thus, cellular images following individual knockdown of these two genes should exhibit phenotypic similarities. In contrast, *SFRP1* (secreted frizzled-related protein 1) forms an inhibitory complex with the frizzled receptor and down-regulates Wnt signaling [22]. Hence, this should show a dissimilar

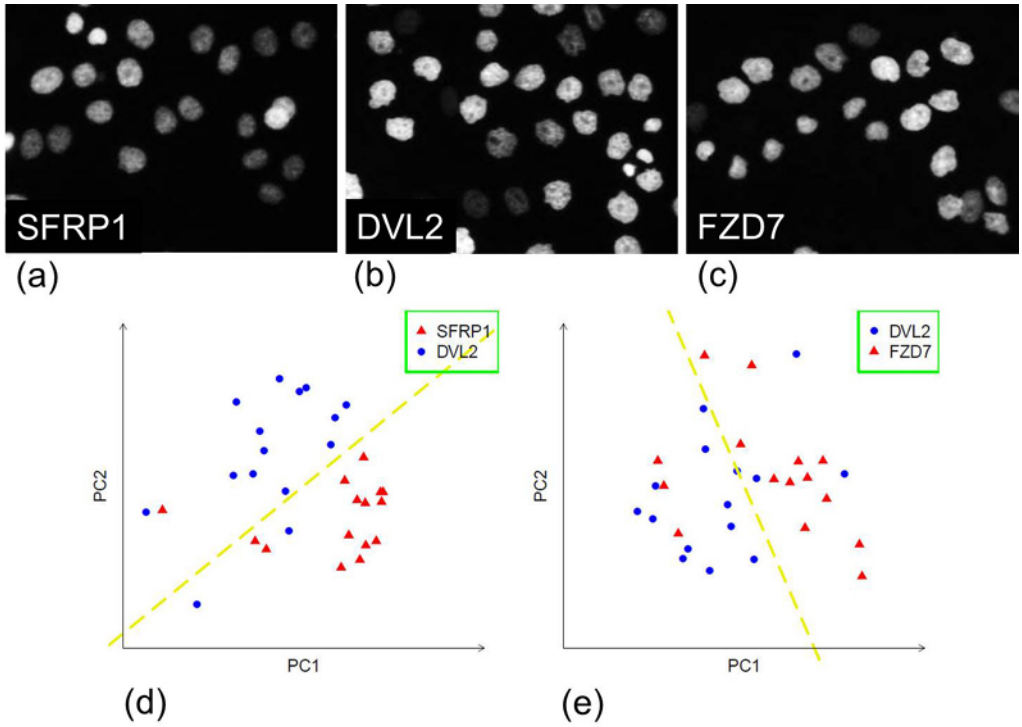

Figure 3.5: Illustration of the characterization of phenotypic similarity. **(a)-(c)** Images of cells in which *SFRP1,DVL2* and *FZD7*, respectively, were knocked down. **(d)** First two principal components (PC 1 and PC 2) of the features for cells in which *SFRP1* and *DVL2* were knocked down; **(e)** first two principal components of the features for cells in which *DVL2* and *FZD7* were knocked down. The dotted lines indicate a linear separation.

cellular phenotype after knockdown. Indeed, after knockdown of *FZD7* and *DVL2*, cells displayed considerably irregular nuclear membranes (Figure 3.5 (b) and Figure 3.5 (c)). In contrast, after knockdown of *SFRP1*, cells did not show these irregular patterns (Figure 3.5 (a)) and were therefore better distinguishable from cells after *FZD7* and *DVL2* knockdown. We segmented the cells in all images and calculated a broad range of texture, morphological and shape features for each cell. Feature vectors were compared for cells in which *SFRP1* and *DVL2* were knocked down (dissimilar images) and in which *DVL2* and *FZD7* were knocked down (similar images). Figure 3.5 (d) and Figure 3.5 (e) show the results from plotting the first two principal components (the first two axes associated with the highest variance of the

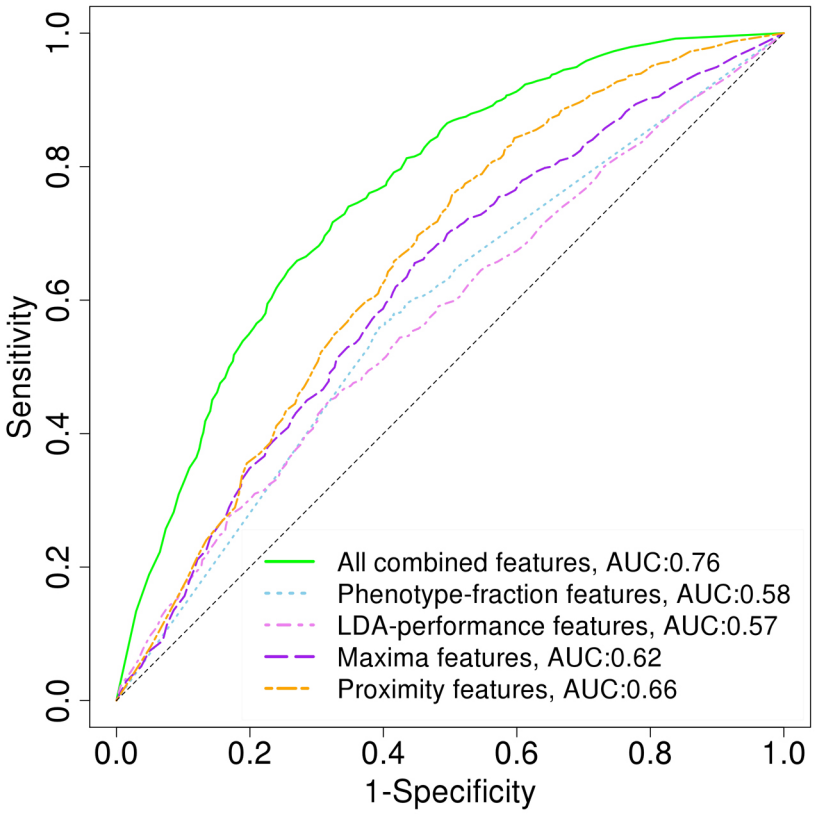

Figure 3.6: ROC curves of the classification using subsets of features. We trained and tested the machines with different subsets of features according to the different types of phenotypic similarity features: the phenotype fraction features, the LDA-performance feature, the maxima features, and the proximity features. The figure shows that the best performance was yielded using all features followed by the set of proximity features.

data in the feature space). With respect to cells in which *FZD7* was knocked down, cells in which *SFRP1* was knocked down were better separable from cells in which *DVL2* was knocked down. Hence, distinguishing knockdown of *SFRP1* and *DVL2* was easier for the discriminator (LDA), and the LDA therefore yielded better performance values (accuracy: 80.7%) in comparison to *DVL2* and *FZD7* (accuracy: 70.6%). The LDA was applied to all combinations of pairs in our data and was used to computed the accuracy, which was one of our similarity features.

3.2.3 Performance of the identification of activating from inhibiting PPIs

We trained 100 Support Vector Machines to distinguish the set of activating PPIs (Act-PPI) from the set of inhibiting PPIs (Inh-PPI). Training and validation were performed through cross-validation. To obtain different levels of stringency, a voting scheme was applied: when a classifier predicted an activating interaction, a positive vote was contributed. Positive votes from all 100 trained SVM-classifiers were summed to yield the predicted interactions and the number of positive votes was used to define stringency. We were particularly interested in classifiers with high stringency. At the highest stringency, remarkably good precision was yielded when selecting interactions that were predicted unequivocally by all classifiers (precision: 92%; accuracy: 35%, sensitivity: 13%, specificity: 97%). Using a minimum 90% stringency yielded a high precision (89%) with a considerably high specificity (87%); the sensitivity was 39%. Figure 3.6 shows the Receiver Operator Characteristics for all features (area under the curve, AUC=0.76) and for the LDA-performance features alone (AUC=0.57) as well as the phenotype-fraction features (AUC=0.58), maxima-features (AUC=0.62) and proximity features (AUC=0.66). We obtained similar results regarding the inhibiting PPIs as true positives and the activating PPIs as true negatives. Using the consistency score (Section 2.2.9) as a filter, we improved the precision of the classification performance considerably. This score expressed the similar or dissimilar nature of two interacting genes with respect to other genes. Consistency scores were calculated and assigned to all of the interaction pairs in our data. To avoid a bias (overfitting) in computing the consistency scores, we did not take the known interactions from KEGG into account in computing these scores. At a high stringency of 80%, the precision was improved from 89% to 94% with 1187 selected activation and 392 selected inhibition interactions. At a middle stringency of 50%, the precision was improved from 84% to 92% with 1177 selected activation and 323 selected inhibition interaction. At a low stringency of 20%, the precision was improved from 81% to 90% with 1137 selected activation and 249 selected inhibition interactions.

3.2.4 Validation with other PPI datasets

To validate the approach using an independent dataset, we compared our predictions with the annotation of known interactions from a well-curated literature-based

database (MetaCore^{TM} version 6.8, www.genego.com). We applied our method to all possible gene pairs and calculated their consistency scores. To avoid overfitting, we did not take known interactions from the investigated KEGG pathways into account. We observed a significant enrichment of gene pairs from the database for predicting activation (P-value = 3.1E-03) and for predicting inhibition (P-value = 4.0E-04). We were interested in how our predictions relate to putative interactions with a high confidence from computationally inferred and not experimentally validated interaction predictions. We used predicted interactions with high confidence scores (scores $\geq$ 900) from the Search Tool for the Retrieval of Interacting Genes (STRING [136]). Interestingly, we found that these interactions showed a significant enrichment (P-value = 4.6E-03). No significant enrichment for inhibiting interactions was yielded. This raised two interesting aspects. The computationally inferred interactions seemed to consist of considerably more activating functions, and our consistency scores may be suitable for predicting new interactions (which was beyond the scope of this study).

3.2.5 Predictions for non-defined PPIs

From the set of undefined PPIs (Undef-PPI) we selected gene pairs with a high number of votes for activation/inhibition and high/low consistency. After discarding interactions found in the literature database (MetaCore, "specific interaction"), 179 new predictions for activation and new 35 predictions for inhibition were yielded. Note that we yielded predictions with good confidence more for activation which is in accordance with the results presented in Section 3.2.3. All 214 predictions can be found in Appendix A.

We then investigated these predictions in greater detail. Commonly, kinases activate their substrates, whereas phosphatases deactivate them. Hence, we performed enrichment tests for these protein groups to validate our predictions and found considerably higher enrichment of kinases and kinase binding proteins in the predictions of activating interactions (for activation, kinase activity P-value = 1.9E-04, and kinase binding P-value = 2.8E-04; for inhibition, kinase activity P-value = 0.03, and no significance was found for kinase binding). In addition, we identified significant enrichments of phosphatase activities and phosphatase-binding proteins *only* in the predictions of inhibiting activations (phosphatase binding P-value = 0.02, pyrophosphatase activity P-value = 0.01, calcium dependent protein serine/threonine

phosphatase activity P-value = 0.007). The results of all of the enrichment tests are presented in Appendix B. We then analyzed the quality of the predicted interactions of the kinases. We compared the potential kinase activities of our predictions for activation with all non-defined interactions (Undef-PPIs). Quantitative mass spectrometry has been employed in obtaining a massive number of site- and context-specific *in vivo* phosphorylation events [15, 23, 81, 84, 85, 93, 138, 140]. Using these data, computational tools have been developed to predict kinase phosphorylation events [78, 83, 96, 159] among which we used one of them which contained the most of our investigated kinases [159]. We found significant enrichment of predicted phosphorylation events in our predictions of kinases interacting with their potential substrates (P-value = 0.015, ratio of our predictions to predicted phosphorylation events: 1.24, other Undef-PPIs: 0.67) and this confirmed our results.

Additionally, we compared our predictions with the literature. Regarding the top twenty predictions for activation, we found two pairs of genes encoding peptides composing the phospholipase C beta (PLC-β) complex, which are therefore positively interacting. PLC-γ2 was predicted to positively interact with HCK and VAV1. This prediction is in accordance with the literature, as HCK was shown to phosphorylate PLC-γ2 in response to activation of cell surface receptors [80], and VAV family proteins positively regulated PLC-γ isoforms downstream of ITAM (immune receptor tyrosine- based active motif) receptors [6, 90, 113, 116, 143, 144]. We found that SRC positively interacts with NFKB and HCK in accordance with an interesting study addressing an epigenetic switch in which constitutively activated SRC activates NFKB, linking inflammatory pathways to oncogenic cell transformation [62]. HCK and SRC are part of the SRC family of kinases (SFKs) and are able to carry out mutual phosphorylation [101]. As evidence of the predictions regarding inhibiting interactions, we found, *e.g.*, SHP1 to negatively regulate KIT receptor tyrosine kinases [86, 106] and CBLB to negatively regulates CRKL signals in response to TCR stimulation [167].

3.2.5.1 The best prediction results were yielded for interactions with cytokine receptors

We were interested in how our predictions were suited for well-defined subgroups of the signaling network. For this, we investigated three major groups: receptors that initiate the signaling processes in the cell, central (highly connected) proteins in the

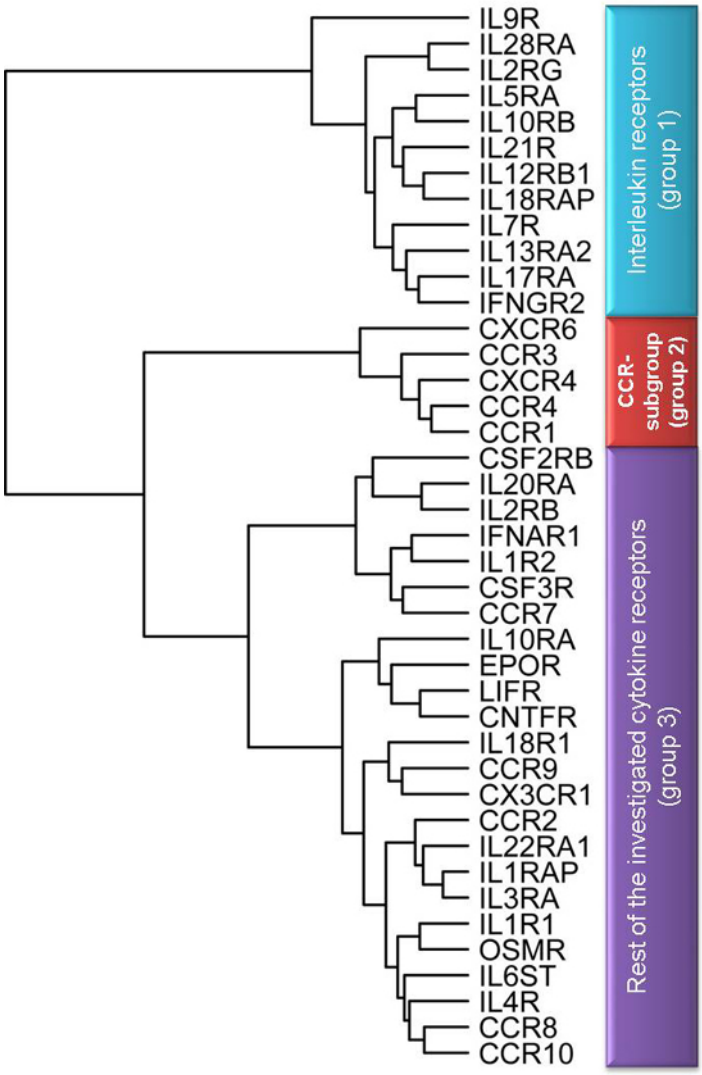

Figure 3.7: Clustering dendrogram for the cytokine receptors. Three groups of clusters were identified: a group of interleukin receptors (group 1), a subgroup of chemokine receptors (group 2) and the rest comprising of interleukin and chemokine receptors.

pathways, and transcription factors as signal destinations. We selected interactions containing at least one node of these groups. A considerably better performance was yielded for the receptors (AUC = 0.92). The group of highly connected proteins showed an average performance (AUC = 0.87), and the transcription factors

performed more poorly (AUC = 0.51), which may reflect their promiscuous functions. The result for the receptors could be improved by restriction to cytokine receptors (AUC = 0.97). In the following section, we describe the investigation this subgroup in more detail.

3.2.5.2 A clustering analysis reveals a new subgroup of chemokine receptors

We employed the consistency scores of the phenotypes and performed unsupervised hierarchical clustering of cytokine receptors. Figure 3.7 shows the results of this analysis. The clustering dendrogram shows three major groups: group 1 mainly consisted of interleukin receptors, group 2 of a subgroup of chemokine receptors (also denoted as the CCR-subgroup hereafter), and group 3 included the remainder of the investigated cytokine receptors. To confirm this clustering, we examined how likely potential interactions within these groups could be formed, employing the information on protein-domain interactions. Interestingly, we found a striking enrichment of domain interactions in the group of chemokine receptors (P-value = 1.8E-09). This was the only group for which any protein was predicted to mutually interact with any other protein in the group (ten mutual interactions). Group 3 showed a much lower, but still significant, enrichment of these interactions (P-value = 0.01, only 42 out of 211 possible interactions). In contrast, the group of interleukins (group 1) showed no enrichment of these protein-domain interactions. Subsequently, we focused our investigations on the detected novel subgroup of five chemokine receptors, *i.e.*, the subgroup of CCR1, CCR3, CCR4, CXCR4 and CXCR6. Clustering all chemokine receptors (using only the consistency scores of the chemokine receptors) confirmed the clustering of the phenotypes of the identified subgroup (Figure 3.8). We further validated that these five CCR genes form a subgroup through a co-expression analysis. We used a large set of 5896 gene expression profiles from microarrays (from 76 different studies from the CAMDA 2007 competition). We compared the correlation of the expression levels of pairs within the subgroup of CCRs with the correlation of pairs within the group of other CCRs. We found a significantly higher correlation in our subgroup of CCRs (P-value = 2.3E-03) demonstrating the close relationship of the CCRs in this subgroup compared to the other CCRs.

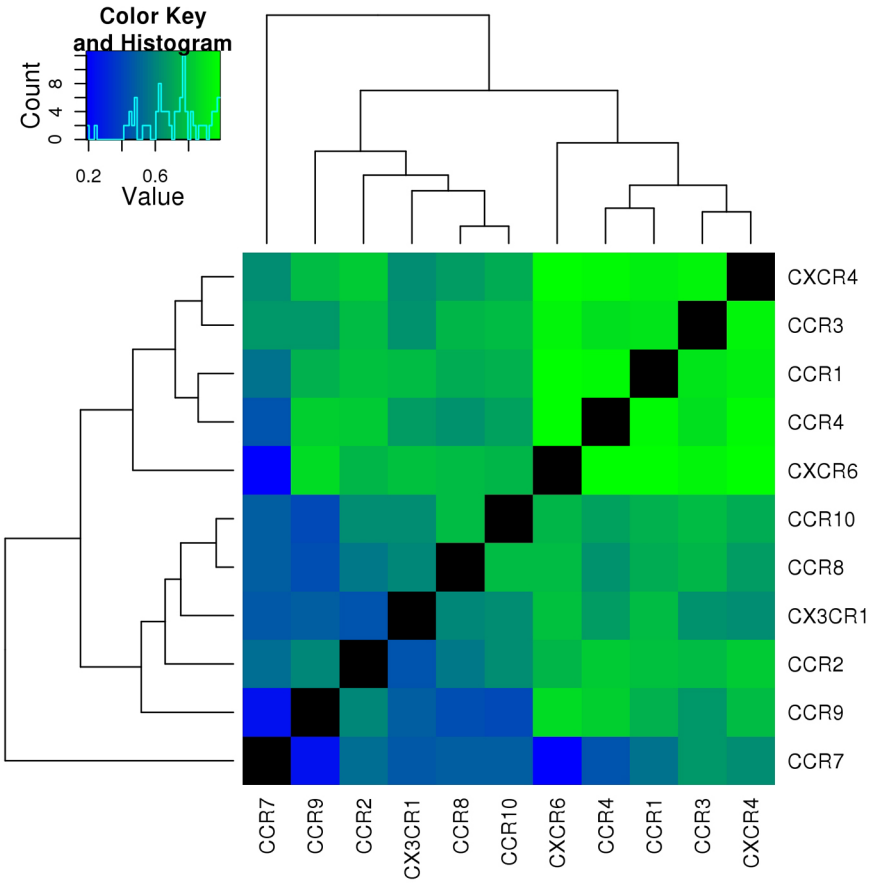

Figure 3.8: Clustering dendrogram for the group of chemokine receptors. The subgroup of chemokine receptors consisting of CCR1, CCR3, CCR4, CXCR4 and CXCR6 clustered together.

3.2.5.3 Investigation of gene groups functionally related to chemokine receptor signaling supports the identification of the new subgroup of CCRs

To elucidate the functional interplay between chemokine receptors and their direct upstream and downstream interactors, we selected the following gene groups: the chemokine receptors themselves; their potentially activating ligands; Jak1, Jak2, Jak3 and Tyk2 as their direct downstream signaling targets activating the Jak/-Stat signaling cascades [127]; G-proteins mediating the PI3kinase/Akt signaling of chemokines (chemokines are G-protein coupled receptors, [157]); and the members

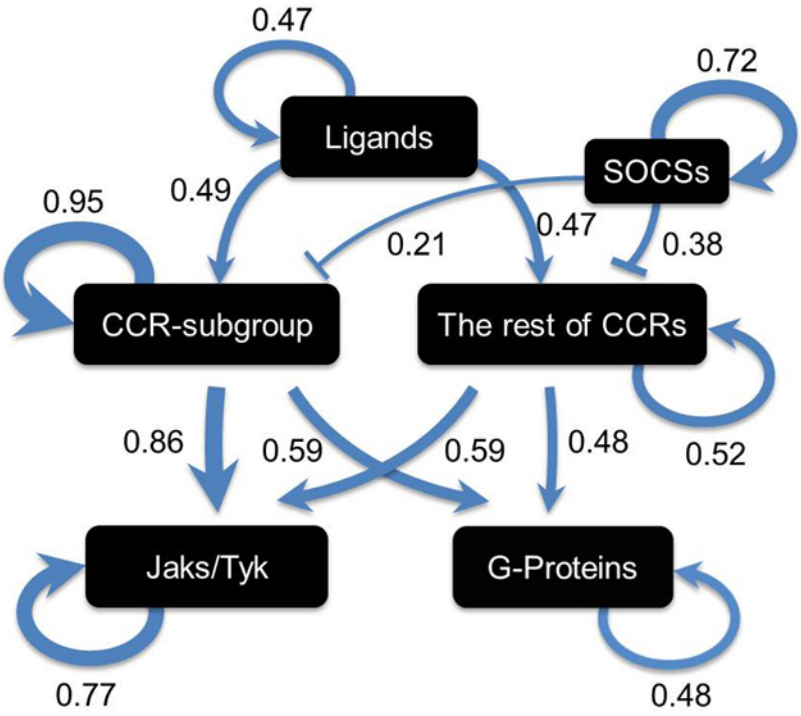

Figure 3.9: Functional interplay of chemokine receptors. This figure shows the functional interplay of chemokine receptors and their direct upstream and downstream interactions. The thickness of the arrows is related to the consistency score between the groups.

of the SOCS family, which inhibit cytokine signaling [162]. To investigate how our subgroup of CCRs is distinguished from the other CCRs, we partitioned the investigated CCRs into two groups: our subgroup and the rest of CCRs. To determine the mutual phenotypic similarity between the investigated groups, we averaged the consistency scores of gene pairs within the groups and between the groups. Figure 3.9 shows the results of this analysis. Thick arrows indicate high consistency, which thick inhibition arcs indicate very inconsistent phenotypes. We found a high consistency within each group (*e.g.*, CCR-subgroup - CCR-subgroup: 0.94, CCR-rest - CCR-rest: 0.59; in comparison, the average consistency of all gene pairs investigated was 0.39). The pairs between all of these groups (except the group of SOCS genes) also showed a higher similarity of phenotypes when compared to the average of all pairs of investigated genes (Table 3.8). As expected, the protein pairs between the

SOCS family member and their target groups showed very low consistency scores, reflecting their inhibitory nature. Interestingly, the pairs including genes from the CCR subgroup exhibited significantly higher consistency scores when compared to the respective pairs containing genes from the group of the rest of CCRs (CCR-subgroup - Ligands vs. CCR-rest - Ligands: P-value = 3.1E-04; CCR-subgroup -Jaks vs. CCR-rest - Jaks: P-value = 1.08E-12; CCR-subgroup - G-protein vs. CCR-rest - G-protein: P-value = 3.41E-07), supporting our hypothesis that they are particularly strongly related.

3.2.5.4 Knockdown of CCR subgroup genes results in higher cell proliferation

We found the CCR subgroup to be distinctively different from the rest of CCRs. We then wanted to obtain insight into how they are different. The maxima features have been well proven to identify and characterize genes related to cell cycle events, such as mitotic delays [98]. We used the maxima features in comparing the genes of the CCR-subgroup with the rest of investigated CCR genes. Interestingly, cells in which genes of the CCR-subgroup were knocked down showed significantly higher

Table 3.8: Mean of consistency score between the CCR-subgroup and other groups.

		Mean of consistency score
CCR-subgroup	Ligands	0.49
CCR-subgroup	Jaks/Tyk	0.86
CCR-subgroup	SOCSs	0.21
CCR-subgroup	G-Proteins	0.59
The rest of CCRs	Ligands	0.47
The rest of CCRs	Jaks/Tyk	0.59
The rest of CCRs	SOCSs	0.38
The rest of CCRs	G-Proteins	0.48
CCR-subgroup	CCR-subgroup	0.95
The rest of CCRs	The rest of CCRs	0.52

proliferation rates compared to when the rest of CCRs were knocked down (P-value = 1.26E-03). Additionally, regarding all investigated CCRs showed that this yielded higher proliferation in comparison to all other investigated genes. Still, for our subgroup, this effect was considerably stronger apparent (significance of higher proliferation of the CCR-subgroup in comparison to all other genes: P-value = 3.73E-06; in contrast, significance of the rest of CCRs versus all investigated genes: P-value = 0.523). We validated these results using an independent public repository produced from a large scale knockdown screen of 72 cell lines from breast, ovarian and pancreas tumors [73]. This screen contained the essentiality information for approximately 16,000 genes, including nearly all (98.6%) of our investigated genes. We argued to have a validation of proliferative influence of a knocked-down gene, if we obtained a depletion of essentiality hits in this screen, when comparing our gene list with the rest of our investigated genes. We again performed two comparisons; we compared 1) our CCR-subgroup with the rest of CCRs, and 2) our CCR-subgroup with the rest of all investigated genes. Both comparisons confirmed the results, comparison CCR-subgroup versus the rest of CCRs: P-value = 4.0E-05, genes from our subgroup were experimentally proven to be 10 times essential and 350 times non-essential in the 72 cell lines (ratio: 0.029), whereas genes from the rest of CCRs were proven to be 39 times essential and 321 times non-essential (ratio: 0.12); comparison CCR-subgroup versus the rest of all investigated genes: P-value = 0.0055, the rest of all investigated genes showed 2894 times to be essential and 44,194 times to be non-essential (ratio: 0.065) in the data of Koh and coworkers [73]. These results confirmed our finding that genes from our CCR subgroup may induce proliferation after knockdown.

Chapter 4

Summary and discussion

4.1 Summary and discussion

In this thesis, I firstly described a new alternative approach to detect host factors (human proteins) involved in HCV infection relying on a fluorescence microscopy imaging of RNAi knockdown screen. The clustering score based on the K-function and was defined to identify the clustering of infected cells. Through the investigation of the alteration of clusters of infected cells after perturbating a gene, we identified a set of potential genes (hits) from our analysis. We compared our targeted host factors with hits from experimental primary and secondary screens, yielding promising gene products that might suit for drug targets. Secondly, I described a new development that based on a machine learning system to characterize the activities of protein-protein interactions from RNAi of HeLa cells. Both published phenotypic descriptors and our own developments were calculated and used for classifications of the activities of protein-protein interaction. A consistency score was established to describe the nature of two interacting proteins and supported to identify the confidence of the predictions. We yielded lists of activation and inhibition predictions. The lists of characterized interactions contributed to our understanding of signal transduction. Additionally, a subset of chemokine receptors was revealed and might yield new insights in chemokine signaling which plays an important role in inflammation and infectious diseases.

4.1.1 Clustering of cells infected with Hepatitis C Virus

We applied a clustering analysis method and statistical analyses of intensity readouts to detect host factors involved in HCV infection. Instead of observing the knockdown of viral components, we focused on specific proteins in the host cell. Targeting host factors that were relevant to viral replication led to distinctively lower clustering of the infected cells. Specifically, all three positive controls showed significantly low clustering scores. Additionally, we obtained hits showing significantly low viral GFP intensities in the primary screen and hits from a secondary screen using a luciferase-based readout. Computation of the intersection of hits from all three approaches yielded five genes to be considered as attractive targets against HCV infection.

Infected cells in the experimental screens showed non-random clustering distributions which has been caused by the spreading of the virus infection. We established a clustering score that was based on the K-function, a distance-based method. The K-function enables to quantify clustering, random, or dispersion distribution at many distances. It allows observing a combination of distributions, *e.g.*, clustering at small scales and regularity at large scales. The combination effects can be observed as a characteristic pattern in a plot of the K-function compared with the null hypothesis determined from a random distribution. Using the K-function, we did not have to pre-define the number of clusters or neighbouring cells before the calculation likes the other methods, as *e.g.*, methods basing on a k-nearest neighbour approach. However, one parameter that we needed to determine for calculating the clustering score was the maximum range of our investigated circular radius from a cell. This parameter depends on the spreading characters of data. We varied the maximum range of the radius and selected the optimal one that showed the optimal correlation coefficients between our clustering score and the intensity values from the experimental screens.

We also applied an alternative clustering method, the quadrat analysis, to measure the infection phenomena of the cell distribution. For this comparison we chose the quadrat analysis method, as similar to the K-function. The quadrat analysis has also the advantage to analyze a distribution statistically in comparison to any distributions. However, the results from our analysis after applying the quadrat analysis showed that the correlation between the clustering score computed from the quadrat analysis and the intensity scores from primary and secondary experimental screens

were much lower than using the K-function. An explanation for this is that the quadrat analysis needs to divide the studied area into a set of grid squares, followed by computing the variability of numbers of points in the grids by a coefficient index that is the variance-mean ratio. If this ratio is nearly to one, the distribution is a random distribution. A ratio greater than one indicates a cluster distribution while a ratio lower than one indicates a uniform distribution. It turned out that the size of the grids is crucial for this. We tried a variety of quadrat sizes. If grid size is too small, there are a lot of empty grid and the statistical test, variance-mean ratio, could not work successfully. If the grid size is too large, it is difficult to detect a cluster distribution of cells. With these limitations, the grid analysis had a bias in detecting the distribution of the cells on our screens.

The homogeneous K-function is commonly used for identifying the distribution of point patterns. The function is computed based on a constantly estimated density for the whole screen. However, for more realistic distributions, the inhomogeneous K-function has been used for our analysis. The inhomogeneous K-function has been used in a wide variety of scientific applications, ranging from the analysis of the clustering behavior of infected habitants in a country [41] to cell biological concerns such as studying the clustering of integrins when cells sense the extra cellular matrix [102]. The inhomogeneous K-function has the same principle as the homogeneous K-function but additionally considers the variation of the intensity over the studied areas. This investigation corresponds to our problem that the infections of cells varied in the different area of the screens which are naturally affected by the spreading of viral particles from a cell to its neighbours. Hence, the inhomogeneous K-function was more suitable for our problem. This explanation was supported by our experimental results which showed the correlation coefficients of the clustering scores for the inhomogeneous K-function and the intensities of the experimental screens were relatively higher than those correlations using the homogeneous K-function. At the maximum radius range of 35% of the images, the correlation coefficients of the clustering scores from the inhomogeneous and homogeneous K-functions with the primary screens were 0.55 and 0.36, respectively. These results supported that the inhomogeneous K-function is more suitable for our application.

We applied established statistical normalization techniques and yielded 30 candidate targets from our clustering methods. When we compared our candidate targets

with the hits from the primary and secondary screens, we yielded five overlapping proteins. With this, we recovered known host factors and new candidate genes. Besides two well-known host factors that are relevant for HCV replication (CD81 and PI4KA) and one host factor that has been described as phosphorylating an HCV protein, we found two new interesting candidates (FLT-4 and SLAMF-6). FLT-4 is a membrane protein and therefore easy targetable by immune cells. It has interesting characteristics. It was observed that it was suited for a retrovirus when genomically incorporated [61]. It will be a challenge to verify FLT-4 experimentally. Then, an important step will be the drug design. To validate the results, gene knockdown experiments may be applied to other liver cell lines that are permissive for HCV replications to observe the infection efficiency of the same knockdown genes.

We used the K-function for observing the clustering behavior of individual infected cells in a cellular *in vitro* assay. With this clustering analysis method, we were able to track the infection of populations in a systematic way and, thus, to detect host factors for viral replication. In addition to apply the K-function to detect relevant host factors as shown in this study, it may be applied to systematically investigate the infection behavior of different virus families. Snijder and co-workers observed principal differences in virus entities to populate cell samples [126]. The K-function may be used in a follow up analysis of the present study through a quantitative clustering analysis supporting a novel taxonomy for virus strains based on their population characteristics in the host. For example, it is known that Dengue virus infects the edges of islets in cell colonies and therefore does not exhibit a clustering tendency like HCV does [126]. In an initial trial, we observed distinct, higher clustering scores for cells infected with HCV in comparison with cells infected by the Dengue virus.

In summary, the application of a clustering analysis method for estimating virulence in cellular assays is general, and this method can be used in other screens to observe infectious propagation in cellular populations. It may also be used to perform a quantitative and systematic analysis of the specific spreading and populating behavior of distinct virus families, which may have an impact on the discovery of their specific use of host cells.

4.1.2 Characterization of signaling interactions

We developed a machine learning-based approach for characterizing the activities of protein-protein interactions. The gold standard of the class labels in the classifications was collected from the KEGG database, which is a comprehensive database providing the interacting information for each interaction with evidences from the literature. A systematic classification was established for distinguishing activation and inhibition interaction using pairwise phenotypic descriptors of gene perturbation. Only a few sets of basic phenotypic similarity features did not yield a good protein-protein activity characterization. In contrast, integrating these features with other developed features yielded a far more comprehensive model. We mainly got our features from four groups comprising the phenotype-fraction features, LDA-performance features, maxima features, and proximity features.

Feature analyses

The performance of the classifiers using each single set of the features were measured. In comparison to the other feature sets, the proximity features yielded the best performance (AUC of 0.66). This makes sense because the proximity features employed the distances computed from both the LDA-performance and the Maxima features. This set of features may contain more variety of information for the classification. Another possible reason might be that the proximity features increased the dimension of the distance features for an interaction. Instead of computing a distance of a protein pair directly, the distance between these two proteins were computed from the distance from each protein to other distinct proteins. The LDA-performance is an important feature even it showed a very low performance (AUC of 0.57) when solely used. When we discarded this LDA-performance feature from our analysis (also discarded from the Proximity features), the overall performance was decreased from the AUC of 0.76 to the AUC of 0.72. The result was similar to the Phenotype-fraction features that the overall performance was decreased to the AUC of 0.73. Therefore, we combined all feature sets for the machine learning system. All feature sets were used for training and testing our machine learning approach using SVMs with a voting scheme technique. An advantage of the technique is the ability to change the stringency parameter to increase precision and to avoid losing potential candidates. The comprehensive model from the machines was used to further predict other interacting protein pairs.

Voting scores and consistency scores

Apart from the voting score for defining the confidence of each predicted interaction, we additionally computed the consistency score for each interacting pair. The consistency score was calculated to identify how likely an interaction processes an activating or inhibiting signal. The consistency score was rigorously computed for an interacting pair, avoiding any biases which could have occurred from the known class labels and trained machines. We used this consistency score to filter out interactions with ambiguous predictions and this improved the precision 7.2% on average. However, a limitation of computing this consistency score is the number of all proteins in the observed pathway. Due to the fact that the score is computed from the percentages of all proteins showing consistent or inconsistent activities with the two proteins for which the prediction is made, higher numbers of proteins yield more reliability. If the number of all proteins in the analysed system is too less, the consistency score might not be appropriate.

Additionally, the performance of the machine learning system to classify the activities of protein-protein interactions to be an activating or inhibiting signal is also important for computing the consistency score. The higher the performance of the classification system to classify the activities is, the more precise these predicted activities become to investigate the consistent (or inconsistent) interaction of a protein pair to other proteins. To obtain a good performance for the classification, the number of samples for training the system is one of the important factors. Generally, many pathways do not provide enough information of the interacting activities. We focused on signal transduction pathways which provided us with high numbers of activating/inhibiting interactions for the machine learning system. The results showed that our systematic classification could be well performed to classify between activating and inhibiting interactions and used in the calculation of the consistency score.

The two scores (from the voting scheme and the consistency score) provided additional information for more precise identification of potential activation and inhibition interactions. Lists of the predicted 179 activations and 35 inhibitions with voting scores and consistency scores were produced. We validated the relevance of our consistency scores using other independent databases through GSEA enrichment tests. The results yielded a significant enrichment of well-defined, known interactions. In the application of the cytokine receptor analysis, we found

a subset of chemokine receptors, consisting of CCR1, CCR3, CCR4, CXCR4 and CXCR6, that showed a significantly high correlation in our co-expression analysis and their relevance is further discussed in the next section.

Application for Cytokine signaling

Cytokine receptors act as dimers or even higher order oligomers [76, 94]. We went into the literature to find evidence of common action of the gene products of the detected subgroup. Seidl *et al.* [122] investigated the gene expression profiles of chemokine receptors using real-time PCR in melanocytes, melanoma cell lines and primary and metastatic melanoma. They found the pair of CXCR4 and CCR1 to be consistently expressed in all of these different melanoma cells, and in the present study, CXCR6 was found to be expressed *de novo* in primary melanomas and melanoma metastases. Among chemokine receptors, CXCR4 and CXCR6 have been reported in several studies to play a predominant role in the development and progression of solid tumors. CXCR4 and CXCR6 interact with tumor cells by activating the AKT/mTor signaling pathway [32]. Furthermore, CXCR4 is also known to activate cancer progression through the JAK/STAT pathway [148], and CXCR4 is associated with a poor prognosis in cervical cancer patients [72]. CXCR4 and CXCR6 are highly expressed in gynecological tumors and in inflammation associated tumors, respectively, and both play important roles in the growth, proliferation, invasion, and metastasis of epithelial ovarian carcinomas [50]. CXCR6 has been found to be highly involved in the metastasis and progression of several types of cancer [32]. The development and aggressiveness of prostate cancer involve the CXCL16/CXCR6 axis [51]. CXCR6 and CXCR4 are expressed in similar proportions in malignant prostate tumors and benign prostate hyperplasia tissue, and both are highly expressed in malignant tissue [60]. Our results demonstrate the common phenotypes of CXCR4 and CXCR6. CCR1, CCR3, CCR4 and CXCR4 were reported to function in human platelets activated in patients infected with human immunodeficiency virus (HIV) and may be commonly involved in inflammatory or allergic responses [24]. Interestingly, in HeLa cells, it was shown that CXCR4 was cross-desensitized by a ligand for CCR4. In chemotaxis, CKLF1 is an activator of CCR4, and SDF1 is an activator of CXCR4. CKLF1 could inhibit the effect of SDF1, which was mediated by CCR4, as SDF1 could be rescued, acting as an activator of chemotaxis after blocking CCR4 [168]. Together with our finding of similar knockdown phenotypes of these receptors, we suggest that both receptors may signal

through very similar downstream cascades. This may be potentiated for one receptor when the other receptor is absent, leading to the same phenotypic shape, regardless of which of the receptors is expressed. Taken together, these results suggest that the member of our subgroup of chemokine receptors exhibit similar functions, and they may even follow similar signaling routes, leading to similar phenotypes following being knockdown.

In conclusion, we developed a machine learning-based approach for predicting interactions involved in the activation or inhibition in the signal transductions. The machines integrated all our developed and published pairwise phenotypic descriptors. We established a consistency score which can be applied to identify the nature of two interacting proteins. Our developed approach is general and can be broadly applied to a larger signaling network. This approach can be exploited to avoid experimental limits of time and cost and might also be applied to the analyses of human disease pathways and networks.

4.2 Outlook

Investigation phenotypes from double knockdowns of infected cells associated with known host factors compared to others may reveal insight with respect to identifying missing cooperative host factors. With this, machine learning can be applied to recognize infected cell characteristics and perform combined host factor predictions. Integrating the cell characterization information and the host-pathogen interaction network may reveal a set of host cofactor proteins that the virus need for the replication. RNAi screens using time-lapse imaging of cells infected with the virus may explore different temporal patterns of infected cells. Tracking the clustering of cells in different time steps may reveal the optimal functional time point for which the virus requires the host factors.

Signaling networks are highly complex. Thus, understanding the system requires a global view of cellular networks. In characterizing the activities of protein-protein interactions, it will be a challenge to integrate all available and reliable protein-protein interaction information from all databases. It will also be a challenge to apply our method to all possible signaling pathways. This application will aid in achieving improved insight into the whole system of a human cell. Validating the

predicted activation and inhibition interactions with protein structures will also be useful to obtain additional evidence of these interactions. In terms of the pairwise phenotypic descriptors, similarities of protein sequences are interesting additional features to explore. Gene ontology may provide hints to identify proteins that are related in terms of function. Additionally, Tan *et al.* [138] reconstructed a conserved phosphorylation (kinase-substrate) network. These phosphorylation events could be included in our model. Furthermore, we can apply our method in a smaller kinome-wide RNAi screen to observe which kinases are activating and inactivating for a given specific treatment or condition.

Appendix A

List of predicted activation and inhibition interactions

Table A.1: List of predicted activation interactions with high vote scores (900-1000) ranked with consistency scores (higher than average); interactions are taken from the STRING database with predicted Domain-interaction database and specific MetaCore interactions are discarded.

Gene ID1	**Gene ID2**	**Gene Name 1**	**Gene Name 2**	**Votes**	**Consistency score**
6714	4790	SRC	NFKB1	1000	0.9950
6714	8440	SRC	NCK2	998	0.9903
6714	10746	SRC	MAP3K2	999	0.9892
23236	5336	PLCB1	PLCG2	989	0.9892
1230	10663	CCR1	CXCR6	1000	0.9834
658	7046	BMPR1B	TGFBR1	975	0.9788
5332	23236	PLCB4	PLCB1	1000	0.9664
1499	7046	CTNNB1	TGFBR1	1000	0.9627
4171	4172	MCM2	MCM3	913	0.9604
6714	3055	SRC	HCK	1000	0.9574
93	658	ACVR2B	BMPR1B	902	0.9534
5970	4790	RELA	NFKB1	958	0.9513
3717	4790	JAK2	NFKB1	995	0.9468
3055	5336	HCK	PLCG2	998	0.9444
6714	5604	SRC	MAP2K1	995	0.9444

Continued on next page...

Table A.1 – continued from previous page

Gene ID1	Gene ID2	Gene Name 1	Gene Name 2	Votes	Consistency score
1232	10663	CCR3	CXCR6	1000	0.9436
5336	7409	PLCG2	VAV1	949	0.9436
658	657	BMPR1B	BMPR1A	965	0.9384
5331	23236	PLCB3	PLCB1	992	0.9375
6714	5582	SRC	PRKCG	987	0.9266
5332	5336	PLCB4	PLCG2	985	0.9233
93	92	ACVR2B	ACVR2A	1000	0.9199
6714	5880	SRC	RAC2	951	0.9151
2066	2065	ERBB4	ERBB3	937	0.9139
6360	10563	CCL16	CXCL13	967	0.9126
2921	10563	CXCL3	CXCL13	999	0.9101
8797	4790	TNFRSF10A	NFKB1	999	0.9047
6352	10563	CCL5	CXCL13	974	0.8931
2921	6366	CXCL3	CCL21	989	0.8900
6366	2919	CCL21	CXCL1	994	0.8836
92	657	ACVR2A	BMPR1A	916	0.8819
6772	4790	STAT1	NFKB1	998	0.8768
939	7186	CD27	TRAF2	1000	0.8714
8440	8976	NCK2	WASL	959	0.8714
6387	10563	CXCL12	CXCL13	966	0.8677
6363	10563	CCL19	CXCL13	914	0.8600
5331	5336	PLCB3	PLCG2	998	0.8520
5608	57551	MAP2K6	TAOK1	996	0.8456
4690	6714	NCK1	SRC	997	0.8412
958	7186	CD40	TRAF2	999	0.8345
6360	6366	CCL16	CCL21	913	0.8345
408	409	ARRB1	ARRB2	994	0.8298
3055	9564	HCK	BCAR1	997	0.8251
6387	6366	CXCL12	CCL21	945	0.8251
6654	6714	SOS1	SRC	1000	0.8226
4792	7186	NFKBIA	TRAF2	995	0.8177
5608	25	MAP2K6	ABL1	974	0.8100
2268	3055	FGR	HCK	966	0.8074
5155	5154	PDGFB	PDGFA	947	0.8074
5584	4790	PRKCI	NFKB1	974	0.8048
6654	8440	SOS1	NCK2	1000	0.7994
1436	9564	CSF1R	BCAR1	996	0.7966

Continued on next page...

Table A.1 – continued from previous page					
Gene ID1	**Gene ID2**	**Gene Name 1**	**Gene Name 2**	**Votes**	**Consistency score**
4772	4775	NFATC1	NFATC3	991	0.7938
1499	5590	CTNNB1	PRKCZ	950	0.7938
2268	6714	FGR	SRC	1000	0.7853
6714	5590	SRC	PRKCZ	994	0.7853
4790	5590	NFKB1	PRKCZ	976	0.7824
998	5608	CDC42	MAP2K6	979	0.7734
6352	5473	CCL5	PPBP	993	0.7673
1237	10663	CCR8	CXCR6	948	0.7641
6367	6376	CCL22	CX3CL1	947	0.7610
3554	4790	IL1R1	NFKB1	923	0.7610
6366	6363	CCL21	CCL19	993	0.7578
6885	659	MAP3K7	BMPR2	995	0.7545
5196	5473	PF4	PPBP	943	0.7545
6714	1398	SRC	CRK	985	0.7512
1237	1232	CCR8	CCR3	967	0.7512
2268	5336	FGR	PLCG2	947	0.7446
3716	5608	JAK1	MAP2K6	998	0.7412
8312	7046	AXIN1	TGFBR1	954	0.7378
2185	5582	PTK2B	PRKCG	939	0.7378
6352	6367	CCL5	CCL22	996	0.7343
3572	3055	IL6ST	HCK	953	0.7343
6654	3055	SOS1	HCK	982	0.7272
1237	1230	CCR8	CCR1	968	0.7237
2921	2919	CXCL3	CXCL1	922	0.7237
6654	5296	SOS1	PIK3R2	1000	0.7200
6352	6360	CCL5	CCL16	999	0.7164
1432	6714	MAPK14	SRC	1000	0.7089
998	5880	CDC42	RAC2	989	0.7051
5321	5604	PLA2G4A	MAP2K1	1000	0.7013
4214	4790	MAP3K1	NFKB1	991	0.7013
1432	2057	MAPK14	EPOR	948	0.7013
92	4092	ACVR2A	SMAD7	998	0.6975
3572	7409	IL6ST	VAV1	960	0.6975
2921	5473	CXCL3	PPBP	942	0.6975
6774	51701	STAT3	NLK	954	0.6936
994	995	CDC25B	CDC25C	919	0.6856
8795	4790	TNFRSF10B	NFKB1	993	0.6816

Continued on next page...

Table A.1 – continued from previous page

Gene ID1	Gene ID2	Gene Name 1	Gene Name 2	Votes	Consistency score
1432	1499	MAPK14	CTNNB1	999	0.6692
9180	3572	OSMR	IL6ST	970	0.6608
6372	5473	CXCL6	PPBP	939	0.6565
4214	5970	MAP3K1	RELA	966	0.6522
5584	998	PRKCI	CDC42	951	0.6434
324	699	APC	BUB1	974	0.6390
2268	9564	FGR	BCAR1	970	0.6390
3627	5196	CXCL10	PF4	964	0.6390
3716	3055	JAK1	HCK	994	0.6345
659	4091	BMPR2	SMAD6	927	0.6345
2064	7046	ERBB2	TGFBR1	984	0.6253
3716	4296	JAK1	MAP3K11	962	0.6253
6654	5880	SOS1	RAC2	994	0.6207
1271	3572	CNTFR	IL6ST	993	0.6207
5604	6776	MAP2K1	STAT5A	990	0.6207
3815	7409	KIT	VAV1	939	0.6113
1432	5604	MAPK14	MAP2K1	925	0.6113
3627	6366	CXCL10	CCL21	923	0.6113
6774	4790	STAT3	NFKB1	992	0.6066
1432	4215	MAPK14	MAP3K3	991	0.6018
6352	6363	CCL5	CCL19	963	0.6018
5580	5604	PRKCD	MAP2K1	960	0.6018
25	9564	ABL1	BCAR1	951	0.6018
3717	4792	JAK2	NFKBIA	995	0.5969
5576	6195	PRKAR2A	RPS6KA1	979	0.5969
3716	5604	JAK1	MAP2K1	980	0.5920
5473	2919	PPBP	CXCL1	925	0.5920
5295	4790	PIK3R1	NFKB1	990	0.5821
6360	6363	CCL16	CCL19	983	0.5821
6352	6372	CCL5	CXCL6	981	0.5821
5295	23236	PIK3R1	PLCB1	956	0.5821
5473	6363	PPBP	CCL19	953	0.5821
6363	2919	CCL19	CXCL1	947	0.5821
5604	7531	MAP2K1	YWHAE	925	0.5771
4214	6885	MAP3K1	MAP3K7	1000	0.5720
6654	2549	SOS1	GAB1	968	0.5720
6352	6361	CCL5	CCL17	907	0.5720

Continued on next page...

Table A.1 – continued from previous page					
Gene ID1	**Gene ID2**	**Gene Name 1**	**Gene Name 2**	**Votes**	**Consistency score**
1432	3055	MAPK14	HCK	949	0.5669
9133	4342	CCNB2	MOS	932	0.5669
4690	5747	NCK1	PTK2	998	0.5618
5335	5336	PLCG1	PLCG2	993	0.5618
8569	5321	MKNK1	PLA2G4A	945	0.5566
1432	3716	MAPK14	JAK1	932	0.5566
2064	5604	ERBB2	MAP2K1	977	0.5513
3627	10563	CXCL10	CXCL13	963	0.5513
4214	4296	MAP3K1	MAP3K11	934	0.5513
5576	5566	PRKAR2A	PRKACA	997	0.5407
10681	5308	GNB5	PITX2	990	0.5407
4914	25	NTRK1	ABL1	981	0.5353
5770	1436	PTPN1	CSF1R	975	0.5353
6654	4690	SOS1	NCK1	967	0.5353
5604	6774	MAP2K1	STAT3	984	0.5299
6370	2919	CCL25	CXCL1	939	0.5299
998	4690	CDC42	NCK1	994	0.5244
5473	6370	PPBP	CCL25	951	0.5244
5576	5577	PRKAR2A	PRKAR2B	987	0.5189
4091	657	SMAD6	BMPR1A	910	0.5134
658	7048	BMPR1B	TGFBR2	964	0.5078
6372	6366	CXCL6	CCL21	920	0.5078
2057	6772	EPOR	STAT1	908	0.5078
1398	5747	CRK	PTK2	903	0.5078
658	4089	BMPR1B	SMAD4	999	0.5022
8408	9706	ULK1	ULK2	984	0.5022
1432	3718	MAPK14	JAK3	973	0.5022
2921	6363	CXCL3	CCL19	970	0.5022
4215	7531	MAP3K3	YWHAE	962	0.4908
3551	8797	IKBKB	TNFRSF10A	943	0.4908
5335	7409	PLCG1	VAV1	910	0.4908
6370	6363	CCL25	CCL19	961	0.4850
207	2065	AKT1	ERBB3	999	0.4792
5335	5332	PLCG1	PLCB4	992	0.4792
2064	5747	ERBB2	PTK2	959	0.4733
5159	5747	PDGFRB	PTK2	985	0.4675
4792	5590	NFKBIA	PRKCZ	910	0.4675

Continued on next page...

Table A.1 – continued from previous page

Gene ID1	Gene ID2	Gene Name 1	Gene Name 2	Votes	Consistency score
5156	5159	PDGFRA	PDGFRB	923	0.4615
994	993	CDC25B	CDC25A	917	0.4615
1031	990	CDKN2C	CDC6	988	0.4555
5568	5577	PRKACG	PRKAR2B	971	0.4555
6361	6367	CCL17	CCL22	951	0.4555
5770	9564	PTPN1	BCAR1	979	0.4495
5921	5894	RASA1	RAF1	949	0.4435
5500	5499	PPP1CB	PPP1CA	933	0.4435
8312	207	AXIN1	AKT1	1000	0.4374
57154	657	SMURF1	BMPR1A	974	0.4374
5563	10891	PRKAA2	PPARGC1A	942	0.4374
2064	3717	ERBB2	JAK2	999	0.4312
5604	5295	MAP2K1	PIK3R1	994	0.4251
815	816	CAMK2A	CAMK2B	990	0.4251
7048	1499	TGFBR2	CTNNB1	993	0.4063
5575	5568	PRKAR1B	PRKACG	983	0.3936

Table A.2: List of predicted inhibition interactions with low vote scores (0-100) ranked with consistency scores (lower than average); interactions are taken from the STRING database with predicted Domain-interaction database and specific MetaCore interactions are discarded.

Gene ID1	**Gene ID2**	**Gene Name 1**	**Gene Name 2**	**Votes**	**Consistency score**
868	1399	CBLB	CRKL	0	-0.3006
891	1022	CCNB1	CDK7	26	-0.2517
868	1398	CBLB	CRK	66	-0.1796
5777	3815	PTPN6	KIT	80	-0.1353
701	57551	BUB1B	TAOK1	20	-0.1130
701	324	BUB1B	APC	19	-0.0980
867	5296	CBL	PIK3R2	1	-0.0151
207	10971	AKT1	YWHAQ	33	-0.0151
701	7157	BUB1B	TP53	88	0.0227
1432	6300	MAPK14	MAPK12	1	0.0755
5970	5595	RELA	MAPK3	60	0.0830
5598	5595	MAPK7	MAPK3	48	0.1353
867	2268	CBL	FGR	5	0.1796
1432	5330	MAPK14	PLCB2	22	0.1869
3265	3667	HRAS	IRS1	27	0.2231
701	890	BUB1B	CCNA2	5	0.2374
3643	5295	INSR	PIK3R1	5	0.2374
701	51343	BUB1B	FZR1	23	0.2374
7132	4217	TNFRSF1A	MAP3K5	21	0.2446
112	114	ADCY6	ADCY8	58	0.2658
112	108	ADCY6	ADCY2	88	0.2658
5601	4804	MAPK9	NGFR	88	0.2728
1432	9021	MAPK14	SOCS3	7	0.2937
5516	5515	PPP2CB	PPP2CA	93	0.2937
5335	5330	PLCG1	PLCB2	1	0.3006
3575	6777	IL7R	STAT5B	22	0.3006
108	196883	ADCY2	ADCY4	45	0.3006
867	4690	CBL	NCK1	5	0.3074
4792	207	NFKBIA	AKT1	98	0.3143
998	56924	CDC42	PAK6	23	0.3479
3554	929	IL1R1	CD14	6	0.3546
5530	5532	PPP3CA	PPP3CB	50	0.3742

Continued on next page...

Table A.2 – continued from previous page

Gene ID1	Gene ID2	Gene Name 1	Gene Name 2	Votes	Consistency score
1871	894	E2F3	CCND2	97	0.3742
115	112	ADCY9	ADCY6	36	0.3807
1019	8900	CDK4	CCNA1	6	0.3872

Appendix B

Gene ontology enrichments

Table B.1: Significant Gene Ontology enrichments of genes in the predicted activation interactions.

GO.ID	**Term**	**Annotated**	**Significant**	***P*-value**
GO:0004713	protein tyrosine kinase activity	152	54	9.50E-07
GO:0042379	chemokine receptor binding	29	17	5.80E-06
GO:0008009	chemokine activity	27	16	9.30E-06
GO:0004674	protein serine/threonine kinase activity	157	52	2.20E-05
GO:0004672	protein kinase activity	171	55	3.20E-05
GO:0004715	non-membrane spanning protein tyrosine kinase activity	14	10	5.40E-05
GO:0030554	adenyl nucleotide binding	188	58	7.50E-05
GO:0032559	adenyl ribonucleotide binding	188	58	7.50E-05
GO:0005057	receptor signaling protein activity	46	21	7.60E-05
GO:0016773	phosphotransferase activity, alcohol group as acceptor	185	57	9.50E-05
GO:0019901	protein kinase binding	80	30	0.00018
GO:0016301	kinase activity	193	58	0.00019
GO:0019899	enzyme binding	147	47	0.00021
GO:0016772	transferase activity, transferring phosphorus-containing groups	194	58	0.00022
GO:0017076	purine nucleotide binding	203	60	0.00024
GO:0032553	ribonucleotide binding	203	60	0.00024
GO:0032555	purine ribonucleotide binding	203	60	0.00024
GO:0000166	nucleotide binding	208	61	0.00026

Continued on next page...

Table B.1 – continued from previous page

GO ID	Term	Annotated	Significant	*P*-value
GO:0005524	ATP binding	182	55	0.00026
GO:0019900	kinase binding	89	32	0.00028
GO:0070411	I-SMAD binding	9	7	0.00037
GO:0016362	activin receptor activity, type II	5	5	0.00037
GO:0048020	CCR chemokine receptor binding	7	6	0.00044
GO:0004871	signal transducer activity	210	60	0.00074
GO:0060089	molecular transducer activity	210	60	0.00074
GO:0035639	purine ribonucleoside triphosphate binding	197	57	0.00076
GO:0019199	transmembrane receptor protein kinase activity	33	15	0.00097
GO:0016740	transferase activity	203	58	0.00098
GO:0031625	ubiquitin protein ligase binding	30	14	0.00105
GO:0001664	G-protein-coupled receptor binding	47	19	0.00119
GO:0005515	protein binding	593	133	0.00123
GO:0004697	protein kinase C activity	4	4	0.00182
GO:0008603	cAMP-dependent protein kinase regulator activity	4	4	0.00182
GO:0048407	platelet-derived growth factor binding	4	4	0.00182
GO:0046332	SMAD binding	19	10	0.00188
GO:0030674	protein binding, bridging	9	6	0.00359
GO:0004675	transmembrane receptor protein serine/threonine kinase activity	12	7	0.0046
GO:0005024	transforming growth factor beta-activated receptor activity	12	7	0.0046
GO:0004702	receptor signaling protein serine/threonine kinase activity	28	12	0.00592
GO:0004712	protein serine/threonine/tyrosine kinase activity	10	6	0.00743
GO:0030234	enzyme regulator activity	89	28	0.00764
GO:0031735	CCR10 chemokine receptor binding	3	3	0.0089
GO:0004435	phosphatidylinositol phospholipase C activity	8	5	0.01198
GO:0004629	phospholipase C activity	8	5	0.01198
GO:0017002	activin receptor activity	8	5	0.01198
GO:0070412	R-SMAD binding	8	5	0.01198
GO:0005488	binding	645	138	0.01779
GO:0019838	growth factor binding	47	16	0.02059
GO:0008047	enzyme activator activity	36	13	0.02192

Continued on next page...

Table B.1 – continued from previous page

GO ID	Term	Annotated	Significant	*P*-value
GO:0042802	identical protein binding	71	22	0.02224
GO:0004708	MAP kinase kinase activity	9	5	0.02248
GO:0005161	platelet-derived growth factor receptor binding	9	5	0.02248
GO:0042169	SH2 domain binding	9	5	0.02248
GO:0004709	MAP kinase kinase kinase activity	12	6	0.02254
GO:0060090	binding, bridging	12	6	0.02254
GO:0005126	cytokine receptor binding	84	25	0.02514
GO:0034713	type I transforming growth factor beta receptor binding	4	3	0.03014
GO:0035254	glutamate receptor binding	4	3	0.03014
GO:0043621	protein self-association	4	3	0.03014
GO:0000975	regulatory region DNA binding	30	11	0.03071
GO:0001067	regulatory region nucleic acid binding	30	11	0.03071
GO:0044212	transcription regulatory region DNA binding	30	11	0.03071
GO:0005096	GTPase activator activity	10	5	0.03752
GO:0008081	phosphoric diester hydrolase activity	10	5	0.03752
GO:0043028	caspase regulator activity	10	5	0.03752
GO:0004716	receptor signaling protein tyrosine kinase activity	7	4	0.0376
GO:0005160	transforming growth factor beta receptor binding	7	4	0.0376
GO:0031434	mitogen-activated protein kinase kinase binding	7	4	0.0376
GO:0048185	activin binding	7	4	0.0376
GO:0003682	chromatin binding	20	8	0.03789
GO:0004706	JUN kinase kinase kinase activity	2	2	0.04321
GO:0005017	platelet-derived growth factor-activated receptor activity	2	2	0.04321
GO:0008093	cytoskeletal adaptor activity	2	2	0.04321
GO:0030159	receptor signaling complex scaffold activity	2	2	0.04321
GO:0030617	transforming growth factor beta receptor, inhibitory cytoplasmic mediator activity	2	2	0.04321
GO:0031730	CCR5 chemokine receptor binding	2	2	0.04321
GO:0031732	CCR7 chemokine receptor binding	2	2	0.04321
GO:0034711	inhibin binding	2	2	0.04321
GO:0048186	inhibin beta-A binding	2	2	0.04321

Continued on next page...

Table B.1 – continued from previous page

GO ID	Term	Annotated	Significant	*P*-value
GO:0048187	inhibin beta-B binding	2	2	0.04321
GO:0070491	repressing transcription factor binding	2	2	0.04321

Table B.2: Significant Gene Ontology enrichments of genes in the predicted inhibition interactions.

GO.ID	**Term**	**Annotated**	**Significant**	***P*-value**
GO:0051219	phosphoprotein binding	15	8	5.60E-06
GO:0045309	protein phosphorylated amino acid binding	9	6	1.70E-05
GO:0001784	phosphotyrosine binding	7	5	6.10E-05
GO:0004707	MAP kinase activity	8	5	0.00015
GO:0019901	protein kinase binding	80	16	0.0003
GO:0004016	adenylate cyclase activity	9	5	0.00032
GO:0009975	cyclase activity	9	5	0.00032
GO:0016849	phosphorus-oxygen lyase activity	9	5	0.00032
GO:0002020	protease binding	6	4	0.00057
GO:0019900	kinase binding	89	16	0.00111
GO:0016829	lyase activity	12	5	0.00167
GO:0008022	protein C-terminus binding	19	6	0.00292
GO:0003824	catalytic activity	337	38	0.00344
GO:0019899	enzyme binding	147	21	0.00373
GO:0030674	protein binding, bridging	9	4	0.00394
GO:0042169	SH2 domain binding	9	4	0.00394
GO:0003924	GTPase activity	27	7	0.00446
GO:0004723	calcium-dependent protein serine/threonine phosphatase activity	2	2	0.00679
GO:0043559	insulin binding	2	2	0.00679
GO:0008294	calcium- and calmodulin-responsive adenylate cyclase activity	6	3	0.00908
GO:0005057	receptor signaling protein activity	46	9	0.00949
GO:0016462	pyrophosphatase activity	39	8	0.01091
GO:0016817	hydrolase activity, acting on acid anhydrides	39	8	0.01091
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	39	8	0.01091
GO:0017111	nucleoside-triphosphatase activity	39	8	0.01091
GO:0060090	binding, bridging	12	4	0.0128
GO:0046934	phosphatidylinositol-4,5-bisphosphate 3-kinase activity	7	3	0.01496
GO:0019902	phosphatase binding	19	5	0.01557
GO:0035014	phosphatidylinositol 3-kinase regulator activity	3	2	0.01927

Continued on next page...

Table B.2 – continued from previous page

GO ID	Term	Annotated	Significant	*P*-value
GO:0004702	receptor signaling protein serine/threonine kinase activity	28	6	0.02239
GO:0004693	cyclin-dependent protein kinase activity	8	3	0.02254
GO:0043548	phosphatidylinositol 3-kinase binding	8	3	0.02254
GO:0016787	hydrolase activity	112	15	0.03029
GO:0004428	inositol or phosphatidylinositol kinase activity	9	3	0.03184
GO:0005158	insulin receptor binding	9	3	0.03184
GO:0035004	phosphatidylinositol 3-kinase activity	9	3	0.03184
GO:0035591	signaling adaptor activity	9	3	0.03184
GO:0035639	purine ribonucleoside triphosphate binding	197	23	0.03188
GO:0032403	protein complex binding	39	7	0.03489
GO:0017046	peptide hormone binding	4	2	0.03649
GO:0051059	NF-kappaB binding	4	2	0.03649
GO:0004722	protein serine/threonine phosphatase activity	10	3	0.04285
GO:0005159	insulin-like growth factor receptor binding	10	3	0.04285
GO:0047485	protein N-terminus binding	10	3	0.04285
GO:0019903	protein phosphatase binding	17	4	0.04493
GO:0017076	purine nucleotide binding	203	23	0.04503
GO:0032553	ribonucleotide binding	203	23	0.04503
GO:0032555	purine ribonucleotide binding	203	23	0.04503
GO:0005524	ATP binding	182	21	0.04778
GO:0016301	kinase activity	193	22	0.04781
GO:0019904	protein domain specific binding	60	9	0.04999

Bibliography

[1] Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock: *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 25(1):25–29, May 2000.

[2] Baddeley, A. and R. Turner: *spatstat: An R package for analyzing spatial point patterns.* J Stat Softw, 12(6):1–42, 2005.

[3] Baddeley, A. J., J. Moller and R. Waagepetersen: *Non- and semi-parametric estimation of interaction in inhomogeneous point patterns.* Statistica Neerlandica, 54(3):329–350, 2000.

[4] Bakal, Chris, John Aach, George Church and Norbert Perrimon: *Quantitative morphological signatures define local signaling networks regulating cell morphology.* Science, 316(5832):1753–1756, 2007.

[5] Berger, K. L., J. D. Cooper, N. S. Heaton, R. Yoon, T. E. Oakland, T. X. Jordan, G. Mateu, A. Grakoui and G. Randall: *Roles for endocytic trafficking and phosphatidylinositol 4-kinase III alpha in hepatitis C virus replication.* Proc Natl Acad Sci U S A, 106(18):7577–82, 2009.

[6] Bertagnolo, V., M. Marchisio, S. Volinia, E. Caramelli and S. Capitani: *Nuclear association of tyrosine-phosphorylated Vav to phospholipase C-gamma1 and phosphoinositide 3-kinase during granulocytic differentiation of HL-60 cells.* FEBS letters, 441(3):480–4, 1998.

[7] Birmingham, A., L. M. Selfors, T. Forster, D. Wrobel, C. J. Kennedy, E. Shanks, J. Santoyo-Lopez, D. J. Dunican, A. Long, D. Kelleher, Q. Smith, R. L. Beijersbergen, P. Ghazal and C. E. Shamu: *Statistical methods for analysis of high-throughput RNA interference screens.* Nat Methods, 6(8):569–75, 2009.

[8] Blight, K. J., J. A. McKeating and C. M. Rice: *Highly permissive cell lines for subgenomic and genomic hepatitis C virus RNA replication.* J Virol, 76(24):13001–14, 2002.

[9] Boerner, K., J. Hermle, C. Sommer, N. P. Brown, B. Knapp, B. Glass, J. Kunkel, G. Torralba, J. Reymann, N. Beil, J. Beneke, R. Pepperkok, R. Schneider, T. Ludwig, M. Hausmann, F. Hamprecht, H. Erfle, L. Kaderali, H. G. Krausslich and M. J. Lehmann: *From experimental setup to bioinformatics: an RNAi screening platform to identify host factors involved in HIV-1 replication.* Biotechnol J, 5(1):39–49, 2009.

[10] Boland, M. V., M. K. Markey and R. F. Murphy: *Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images.* Cytometry, 33(3):366–75, 1998.

[11] Borawski, J., P. Troke, X. Puyang, V. Gibaja, S. Zhao, C. Mickanin, J. Leighton-Davies, C. J. Wilson, V. Myer, I. Cornellataracido, J. Baryza, J. Tallarico, G. Joberty, M. Bantscheff, M. Schirle, T. Bouwmeester, J. E. Mathy, K. Lin, T. Compton, M. Labow, B. Wiedmann and L. A. Gaither: *Class III phosphatidylinositol 4-kinase alpha and beta are novel host factor regulators of hepatitis C virus replication.* J Virol, 83(19):10058–74, 2009.

[12] Boutros, M., L. P. Bras and W. Huber: *Analysis of cell-based RNAi screens.* Genome Biol, 7(7):R66, 2006.

[13] Brideau, C., B. Gunter, B. Pikounis and A. Liaw: *Improved statistical methods for hit selection in high-throughput screening.* J Biomol Screen, 8(6):634–47, 2003.

[14] Brown, James R., Michal Magid-Slav, Philippe Sanseau and Deepak K. Rajpal: *Computational biology approaches for selecting host-pathogen drug targets.* Drug Discov Today, 16(5-6):229–236, Mar 2011.

[15] Brugge, J. S., G. Jarosik, J. Andersen, A. Queral-Lustig, M. Fedor-Chaiken and J. R. Broach: *Expression of Rous sarcoma virus transforming protein pp60v-src in Saccharomyces cerevisiae cells.* Molecular and cellular biology, 7(6):2180–7, 1987.

[16] Burges, C. J. C.: *A tutorial on Support Vector Machines for pattern recognition.* Data Min Knowl Disc, 2(2):121–167, 1998.

[17] Burkhardt, H. and S. Siggelkow: *Invariant Features in Pattern Recognition – Fundamentals and Applications.* In Kotropoulos, C. and I. Pitas (editors): *Nonlinear Model-Based Image/Video Processing and Analysis*, pages 269–307. John Wiley & Sons, 2001.

[18] Carpenter, A. E., T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland and D. M. Sabatini: *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biol, 7(10):R100–, 2006.

[19] Carter, C. A. and L. S. Ehrlich: *Cell biology of HIV-1 infection of macrophages.* Annu Rev Microbiol, 62:425–43, 2008.

[20] Chang, C.-C. and C.-J. Lin.: *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2(27):1–27, 2011.

[21] Chang, T. and C. J. Kuo: *Texture analysis and classification with tree-structured wavelet transform.* IEEE Trans Image Process, 2(4):429–41, 1993.

[22] Chim, C. S., R. Pang, T. K. Fung, C. L. Choi and R. Liang: *Epigenetic dysregulation of Wnt signaling pathway in multiple myeloma.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 21(12):2527–36, 2007.

[23] Choi, H., B. Larsen, Z. Y. Lin, A. Breitkreutz, D. Mellacheruvu, D. Fermin, Z. S. Qin, M. Tyers, A. C. Gingras and A. I. Nesvizhskii: *SAINT: probabilistic scoring of affinity purification-mass spectrometry data.* Nature methods, 8(1):70–3, 2011.

[24] Clemetson, K. J., J. M. Clemetson, A. E. Proudfoot, C. A. Power, M. Baggiolini and T. N. Wells: *Functional expression of CCR1, CCR3, CCR4, and CXCR4 chemokine receptors on human platelets.* Blood, 96(13):4046–54, 2000.

[25] Cleveland, W. S.: *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 74(368):829–836, 1979.

[26] Cohen, P.: *The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture.* Eur J Biochem, 268(19):5001–10, 2001.

[27] Conrad, C., H. Erfle, P. Warnat, N. Daigle, T. Lorch, J. Ellenberg, R. Pepperkok and R. Eils: *Automatic identification of subcellular phenotypes on human cell arrays.* Genome Res, 14(6):1130–6, 2004.

[28] Conrad, C. and D. W. Gerlich: *Automated microscopy for high-content RNAi screening.* J Cell Biol, 188(4):453–61, 2010.

[29] Conrad, C., A. Wunsche, T. H. Tan, J. Bulkescher, F. Sieckmann, F. Verissimo, A. Edelstein, T. Walter, U. Liebel, R. Pepperkok and J. Ellenberg: *Micropilot: automation of fluorescence microscopy-based imaging for systems biology.* Nat Methods, 8(3):246–9, 2011.

[30] Daubechies, I.: *Orthonormal Bases of Compactly Supported Wavelets.* Communications on Pure and Applied Mathematics, 41(7):909–996, 1988.

[31] Davis, Fred P., David T. Barkan, Narayanan Eswar, James H. McKerrow and Andrej Sali: *Host pathogen protein interactions predicted by comparative modeling.* Protein Sci, 16(12):2585–2596, Dec 2007.

[32] Deng, L., N. Chen, Y. Li, H. Zheng and Q. Lei: *CXCR6/CXCL16 functions as a regulator in metastasis and progression of cancer.* Biochimica et biophysica acta, 1806(1):42–9, 2010.

[33] Doolittle, Janet M. and Shawn M. Gomez: *Mapping protein interactions between Dengue virus and its human and insect hosts.* PLoS Negl Trop Dis, 5(2):e954, 2011.

[34] Dubuisson, J., F. Helle and L. Cocquerel: *Early steps of the hepatitis C virus life cycle.* Cell Microbiol, 10(4):821–7, 2008.

[35] Duda, Richard O., Peter E. Hart and David G. Stork: *Pattern Classification.* Wiley-Interscience, New York, 2. edition, 2001.

[36] Dyer, Matthew D., T. M. Murali and Bruno W. Sobral: *Computational prediction of host-pathogen protein-protein interactions.* Bioinformatics, 23(13):i159–i166, Jul 2007.

[37] Echeverri, C. J. and N. Perrimon: *High-throughput RNAi screening in cultured cells: a user's guide.* Nature reviews. Genetics, 7(5):373–84, 2006.

[38] Engel, P., M. J. Eck and C. Terhorst: *The SAP and SLAM families in immune responses and X-linked lymphoproliferative disease.* Nature reviews. Immunology, 3(10):813–21, 2003.

[39] Erfle, H., B. Neumann, U. Liebel, P. Rogers, M. Held, T. Walter, J. Ellenberg and R. Pepperkok: *Reverse transfection on cell arrays for high content screening microscopy.* Nat Protoc, 2(2):392–9, 2007.

[40] Erfle, H., B. Neumann, P. Rogers, J. Bulkescher, J. Ellenberg and R. Pepperkok: *Work flow for multiplexing siRNA assays by solid-phase reverse transfection in multiwell plates.* J Biomol Screen, 13(7):575–80, 2008.

[41] Ersboll, A. K. and B. K. Ersboll: *Simulation of the K-function in the analysis of spatial clustering for non-randomly distributed locations–exemplified by bovine virus diarrhoea virus (BVDV) infection in Denmark.* Prev Vet Med, 91(1):64–71, 2009.

[42] Evans, M. J., T. von Hahn, D. M. Tscherne, A. J. Syder, M. Panis, B. Wolk, T. Hatziioannou, J. A. McKeating, P. D. Bieniasz and C. M. Rice: *Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry.* Nature, 446(7137):801–5, 2007.

[43] Fisher, Ronald A.: *The use of multiple measurements in taxonomic problems.* Annals Eugen., 7:179–188, 1936.

[44] Forestier, N. and S. Zeuzem: *Telaprevir for the treatment of hepatitis C.* Expert Opin Pharmacother, 13(4):593–606, 2012.

[45] Fuchs, Florian, Gregoire Pau, Dominique Kranz, Oleg Sklyar, Christoph Budjan, Sandra Steinbrink, Thomas Horn, Angelika Pedal, Wolfgang Huber and Michael Boutros: *Clustering phenotype populations by genome-wide RNAi and multiparametric imaging.* Mol Syst Biol, 6:370, 2010.

[46] Gambe, Arni E., Rika Maniwa Ono, Sachihiro Matsunaga, Natsumaro Kutsuna, Takumi Higaki, Tsunehito Higashi, Seiichiro Hasezawa, Susumu Uchiyama and Kiichi Fukui: *Development of a multistage classifier for a monitoring system of cell activity based on imaging of chromosomal dynamics.* Cytometry A, 71(5):286–296, May 2007.

[47] Gipp, Markus, Guillermo Marcus, Nathalie Harder, Apichat Suratanee, Karl Rohr, Rainer Knig and Reinhard Mnner: *Haralick's Texture Features Computed by GPUs for Biological Applications.* IAENG International Journal of Computer Science, 36(11):66–75, 2009.

[48] Goshima, Gohta, Roy Wollman, Sarah S. Goodwin, Nan Zhang, Jonathan M. Scholey, Ronald D. Vale and Nico Stuurman: *Genes required for mitotic spindle assembly in Drosophila S2 cells.* Science, 316(5823):417–421, Apr 2007.

[49] Griewank, K., C. Borowski, S. Rietdijk, N. Wang, A. Julien, D. G. Wei, A. A. Mamchak, C. Terhorst and A. Bendelac: *Homotypic interactions mediated by Slamf1 and Slamf6 receptors control NKT cell lineage development.* Immunity, 27(5):751–62, 2007.

[50] Guo, Li, Zhu-Mei Cui, Jia Zhang and Yu Huang: *Chemokine axes CXCL12/CXCR4 and CXCL16/CXCR6 correlate with lymph node metastasis in epithelial ovarian carcinoma.* Chin J Cancer, 30(5):336–343, 2011.

[51] Ha, Hong Koo, Wan Lee, Hyun Jun Park, Sang Don Lee, Jeong Zoo Lee and Moon Kee Chung: *Clinical significance of CXCL16/CXCR6 expression in patients with prostate cancer.* Mol Med Report, 4(3):419–424, 2011.

[52] Harada, Josephine N., Kristen E. Bower, Anthony P. Orth, Scott Callaway, Christian G. Nelson, Casey Laris, John B. Hogenesch, Peter K. Vogt and Sumit K. Chanda: *Identification of novel mammalian growth regulatory factors by genome-scale quantitative image analysis.* Genome Res, 15(8):1136–1144, Aug 2005.

[53] Haralick, R. M.: *Statistical and structural approaches to texture.* Proceedings of the IEEE, 67(5):786–804, May 1979.

[54] Harder, N.: *Automatic Cell Cycle Analysis Based on Live Cell Fluorescence Microscopy Image Sequences.* Logos Verlag Berlin, 2010.

[55] Harder, N., R. Eils and K. Rohr: *Automated classification of mitotic phenotypes of human cells using fluorescent proteins.* Methods Cell Biol, 85:539–554, 2008.

[56] Harder, N., F. Mora-Bermudez, W. J. Godinez, J. Ellenberg, R. Eils and K. Rohr: *Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences.* 9th International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI'06), 9(Pt 1):840–8, 2006.

[57] Harder, N., F. Mora-Bermudez, W. J. Godinez, A. Wunsche, R. Eils, J. Ellenberg and K. Rohr: *Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time.* Genome Res, 19(11):2113–24, 2009.

[58] Held, Michael, Michael H A. Schmitz, Bernd Fischer, Thomas Walter, Beate Neumann, Michael H. Olma, Matthias Peter, Jan Ellenberg and Daniel W. Gerlich: *CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging.* Nat Methods, 7(9):747–754, Sep 2010.

[59] Hsu, Chih-Wei, Chih-Chung Chang and Chih-Jen Lin: *A Practical Guide to Support Vector Classification*. http://www.csie.ntu.edu.tw/ cjlin/, pages –, 2010.

[60] Hu, W., X. Zhen, B. Xiong, B. Wang, W. Zhang and W. Zhou: *CXCR6 is expressed in human prostate cancer in vivo and is involved in the in vitro invasion of PC3 and LNCap cells*. Cancer science, 99(7):1362–9, 2008.

[61] Hughes, D. C.: *Alternative splicing of the human VEGFGR-3/FLT4 gene as a consequence of an integrated human endogenous retrovirus*. J Mol Evol, 53(2):77–9, 2001.

[62] Iliopoulos, D., H. A. Hirsch and K. Struhl: *An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation*. Cell, 139(4):693–706, 2009.

[63] Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork and C. von Mering: *STRING 8–a global view on proteins and their functional interactions in 630 organisms*. Nucleic Acids Res, 37(Database issue):D412–6, 2009.

[64] Jones, T. R., I. H. Kang, D. B. Wheeler, R. A. Lindquist, A. Papallo, D. M. Sabatini, P. Golland and A. E. Carpenter: *CellProfiler Analyst: data exploration and analysis software for complex image-based screens*. BMC Bioinformatics, 9:482, 2008.

[65] Joyce, M. A. and D. L. Tyrrell: *The cell biology of hepatitis C virus*. Microbes Infect, 12(4):263–71, 2010.

[66] Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa: *From genomics to chemical genomics: new developments in KEGG*. Nucleic acids research, 34(Database issue):D354–7, 2006.

[67] Kanehisa, M., S. Goto, S. Kawashima and A. Nakaya: *The KEGG databases at GenomeNet*. Nucleic acids research, 30(1):42–6, 2002.

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

[69] Keshava Prasad, T. S., Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady and Akhilesh Pandey: *Human Protein Reference Database–2009 update.* Nucleic Acids Res, 37(Database issue):D767–D772, Jan 2009.

[70] Kim, J., D. Lee and J. Choe: *Hepatitis C virus NS5A protein is phosphorylated by casein kinase II.* Biochem Biophys Res Commun, 257(3):777–81, 1999.

[71] Kitano, Hiroaki: *Biological robustness in complex host-pathogen systems.* Prog Drug Res, 64:239, 241–239, 263, 2007.

[72] Kodama, J., Hasengaowa, T. Kusumoto, N. Seki, T. Matsuo, Y. Ojima, K. Nakamura, A. Hongo and Y. Hiramatsu: *Association of CXCR4 and CCR7 chemokine receptor expression and lymph node metastasis in human cervical cancer.* Ann Oncol, 18(1):70–76, 2007.

[73] Koh, J. L., K. R. Brown, A. Sayad, D. Kasimer, T. Ketela and J. Moffat: *COLT-Cancer: functional genetic screening resource for essen-*

tial genes in human cancer cell lines. Nucleic acids research, 40(Database issue):D957–63, 2012.

[74] Lage, Kasper, Kjeld Mllgrd, Steven Greenway, Hiroko Wakimoto, Joshua M. Gorham, Christopher T. Workman, Eske Bendsen, Niclas T. Hansen, Olga Rigina, Francisco S. Roque, Cornelia Wiese, Vincent M. Christoffels, Amy E. Roberts, Leslie B. Smoot, William T. Pu, Patricia K. Donahoe, Niels Tommerup, Sren Brunak, Christine E. Seidman, Jonathan G. Seidman and Lars A. Larsen: *Dissecting spatio-temporal protein networks driving human heart development and related disorders.* Mol Syst Biol, 6:381, 2010.

[75] Lamprecht, M. R., D. M. Sabatini and A. E. Carpenter: *CellProfiler: free, versatile software for automated biological image analysis.* Biotechniques, 42(1):71–5, 2007.

[76] Lemmon, Mark A. and Joseph Schlessinger: *Cell signaling by receptor tyrosine kinases.* Cell, 141(7):1117–1134, 2010.

[77] Li, Q., A. L. Brass, A. Ng, Z. Hu, R. J. Xavier, T. J. Liang and S. J. Elledge: *A genome-wide genetic screen for host factors required for hepatitis C virus propagation.* Proc Natl Acad Sci U S A, 106(38):16410–5, 2009.

[78] Li, T., F. Li and X. Zhang: *Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach.* Proteins, 70(2):404–14, 2008.

[79] Lian, Z., J. Liu, M. Wu, H. Y. Wang, P. Arbuthnot, M. Kew and M. A. Feitelson: *Hepatitis B x antigen up-regulates vascular endothelial growth factor receptor 3 in hepatocarcinogenesis.* Hepatology, 45(6):1390–9, 2007.

[80] Liao, F., H. S. Shin and S. G. Rhee: *In vitro tyrosine phosphorylation of PLC-gamma 1 and PLC-gamma 2 by src-family protein tyrosine kinases.* Biochemical and biophysical research communications, 191(3):1028–33, 1993.

[81] Lim, W. A. and T. Pawson: *Phosphotyrosine signaling: evolving a new cellular communication system.* Cell, 142(5):661–7, 2010.

[82] Lindblad, Joakim, Carolina Whlby, Ewert Bengtsson and Alla Zaltsman: *Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation.* Cytometry A, 57(1):22–33, Jan 2004.

[83] Linding, R., L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jorgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe and T. Pawson: *Systematic discovery of in vivo phosphorylation networks.* Cell, 129(7):1415–26, 2007.

[84] Liu, B. A., K. Jablonowski, E. E. Shah, B. W. Engelmann, R. B. Jones and P. D. Nash: *SH2 domains recognize contextual peptide sequence information to determine selectivity.* Molecular & cellular proteomics : MCP, 9(11):2391–404, 2010.

[85] Liu, G., J. Zhang, B. Larsen, C. Stark, A. Breitkreutz, Z. Y. Lin, B. J. Breitkreutz, Y. Ding, K. Colwill, A. Pasculescu, T. Pawson, J. L. Wrana, A. I. Nesvizhskii, B. Raught, M. Tyers and A. C. Gingras: *ProHits: integrated software for mass spectrometry-based interaction proteomics.* Nature biotechnology, 28(10):1015–7, 2010.

[86] Lorenz, U., A. D. Bergemann, H. N. Steinberg, J. G. Flanagan, X. Li, S. J. Galli and B. G. Neel: *Genetic analysis reveals cell type-specific regulation of receptor tyrosine kinase c-Kit by the protein tyrosine phosphatase SHP1.* The Journal of experimental medicine, 184(3):1111–26, 1996.

[87] M. Roula, A. Bouridan, f. Kurugollu J. Diamond: *3D segmentation and feature extraction of CLSM scanned nuclei using evolutionary snakes.* In *Proc. IEEE Internat. Symposium on Biomedical Imaging: From Nano to Macro (ISBI'2007)*, pages 316–319, 2007.

[88] Malim, M. H. and M. Emerman: *HIV-1 accessory proteins–ensuring viral survival in a hostile environment.* Cell Host Microbe, 3(6):388–98, 2008.

[89] Malo, N., J. A. Hanley, S. Cerquozzi, J. Pelletier and R. Nadon: *Statistical practice in high-throughput screening data analysis.* Nature biotechnology, 24(2):167–75, 2006.

[90] Manetz, T. S., C. Gonzalez-Espinosa, R. Arudchandran, S. Xirasagar, V. Tybulewicz and J. Rivera: *Vav1 regulates phospholipase cgamma activation and calcium responses in mast cells.* Molecular and cellular biology, 21(11):3763–74, 2001.

[91] Martin, N. and Q. Sattentau: *Cell-to-cell HIV-1 spread and its implications for immune evasion.* Curr Opin HIV AIDS, 4(2):143–9, 2009.

[92] Matula, P., A. Kumar, I. Worz, H. Erfle, R. Bartenschlager, R. Eils and K. Rohr: *Single-cell-based image analysis of high-throughput cell array screens for quantification of viral infection.* Cytometry A, 75(4):309–18, 2009.

[93] Miller, M. L., L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak and R. Linding: *Linear motif atlas for phosphorylation-dependent signaling.* Science signaling, 1(35):ra2, 2008.

[94] Milligan, Graeme: *G protein-coupled receptor dimerisation: molecular basis and relevance to function.* Biochim Biophys Acta, 1768(4):825–835, 2007.

[95] Moriishi, K. and Y. Matsuura: *Exploitation of lipid components by viral and host proteins for hepatitis C virus infection.* Front Microbiol, 3:54, 2012.

[96] Neuberger, G., G. Schneider and F. Eisenhaber: *pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model.* Biology direct, 2:1, 2007.

[97] Neumann, B., M. Held, U. Liebel, H. Erfle, P. Rogers, R. Pepperkok and J. Ellenberg: *High-throughput RNAi screening by time-lapse imaging of live human cells.* Nat Methods, 3(5):385–90, 2006.

[98] Neumann, B., T. Walter, J. K. Heriche, J. Bulkescher, H. Erfle, C. Conrad, P. Rogers, I. Poser, M. Held, U. Liebel, C. Cetin, F. Sieckmann, G. Pau, R. Kabbe, A. Wunsche, V. Satagopam, M. H.

Schmitz, C. Chapuis, D. W. Gerlich, R. Schneider, R. Eils, W. Huber, J. M. Peters, A. A. Hyman, R. Durbin, R. Pepperkok and J. Ellenberg: *Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes.* Nature, 464(7289):721–7, 2010.

[99] Ng, Teresa I., Hongmei Mo, Tami Pilot-Matias, Yupeng He, Gennadiy Koev, Preethi Krishnan, Rubina Mondal, Ron Pithawalla, Wenping He, Tanya Dekhtyar, Jeremy Packer, Mark Schurdak and Akhteruzzaman Molla: *Identification of host genes involved in hepatitis C virus replication by small interfering RNA technology.* Hepatology, 45(6):1413–1421, Jun 2007.

[100] Otsu, N.: *A threshold selection method from grey level histograms.* IEEE Transactions on Systems, Man and Cybernetics, 9:62–66, 1979.

[101] Page, T. H., M. Smolinska, J. Gillespie, A. M. Urbaniak and B. M. Foxwell: *Tyrosine kinases and inflammatory signalling.* Current molecular medicine, 9(1):69–85, 2009.

[102] Paszek, M. J., D. Boettiger, V. M. Weaver and D. A. Hammer: *Integrin clustering is driven by mechanical resistance from the glycocalyx and the substrate.* PLoS Comput Biol, 5(12):e1000604, 2009.

[103] Pau, G., F. Fuchs, O. Sklyar, M. Boutros and W. Huber: *EBImage–an R package for image processing with applications to cellular phenotypes.* Bioinformatics, 26(7):979–81, 2010.

[104] Peng, T., G. M. Bonamy, E. Glory-Afshar, D. R. Rines, S. K. Chanda and R. F. Murphy: *Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns.* Proceedings of the National Academy of Sciences of the United States of America, 107(7):2944–9, 2010.

[105] Perlman, Zachary E., Timothy J. Mitchison and Thomas U. Mayer: *High-content screening and profiling of drug activity in an automated centrosome-duplication assay.* Chembiochem, 6(1):145–151, Jan 2005.

[106] Piao, X., R. Paulson, P. van der Geer, T. Pawson and A. Bernstein: *Oncogenic mutation in the Kit receptor tyrosine kinase alters substrate specificity and induces degradation of the protein tyrosine phosphatase SHP-1.* Proceedings of the National Academy of Sciences of the United States of America, 93(25):14665–9, 1996.

[107] Pietschmann, T.: *Virology: Final entry key for hepatitis C.* Nature, 457(7231):797–8, 2009.

[108] Pileri, P., Y. Uematsu, S. Campagnoli, G. Galli, F. Falugi, R. Petracca, A. J. Weiner, M. Houghton, D. Rosa, G. Grandi and S. Abrignani: *Binding of hepatitis C virus to CD81.* Science, 282(5390):938–41, 1998.

[109] Pol, S., A. Vallet-Pichard, M. Corouge and V. O. Mallet: *Hepatitis C: epidemiology, diagnosis, natural history and therapy.* Contrib Nephrol, 176:1–9, 2012.

[110] Prior, I. A., C. Muncke, R. G. Parton and J. F. Hancock: *Direct visualization of Ras proteins in spatially distinct cell surface microdomains.* J Cell Biol, 160(2):165–70, 2003.

[111] Prokop, R. J. and A. P. Reeves: *A Survey of Moment-Based Techniques for Unoccluded Object Representation and Recognition.* Cvgip-Graphical Models and Image Processing, 54(5):438–460, 1992.

[112] Racanelli, Vito and Barbara Rehermann: *Hepatitis C virus infection: when silence is deception.* Trends Immunol, 24(8):456–464, Aug 2003.

[113] Ramos-Morales, F., B. J. Druker and S. Fischer: *Vav binds to several SH2/SH3 containing proteins in activated lymphocytes.* Oncogene, 9(7):1917–23, 1994.

[114] Randall, G., M. Panis, J. D. Cooper, T. L. Tellinghuisen, K. E. Sukhodolets, S. Pfeffer, M. Landthaler, P. Landgraf, S. Kan, B. D. Lindenbach, M. Chien, D. B. Weir, J. J. Russo, J. Ju, M. J. Brownstein, R. Sheridan, C. Sander, M. Zavolan, T. Tuschl and

C. M. Rice: *Cellular cofactors affecting hepatitis C virus infection and replication.* Proc Natl Acad Sci U S A, 104(31):12884–9, 2007.

[115] Reiss, Simon, Ilka Rebhan, Perdita Backes, Ines Romero-Brey, Holger Erfle, Petr Matula, Lars Kaderali, Marion Poenisch, Hagen Blankenburg, Marie-Sophie Hiet, Thomas Longerich, Sarah Diehl, Fidel Ramirez, Tamas Balla, Karl Rohr, Artur Kaul, Sandra Buehler, Rainer Pepperkok, Thomas Lengauer, Mario Albrecht, Roland Eils, Peter Schirmacher, Volker Lohmann and Ralf Bartenschlager: *Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment.* Cell Host Microbe, 9(1):32–45, Jan 2011.

[116] Reynolds, L. F., L. A. Smyth, T. Norton, N. Freshney, J. Downward, D. Kioussis and V. L. Tybulewicz: *Vav1 transduces T cell receptor signals to the activation of phospholipase C-gamma1 via phosphoinositide 3-kinase-dependent and -independent pathways.* The Journal of experimental medicine, 195(9):1103–14, 2002.

[117] Rieber, N., B. Knapp, R. Eils and L. Kaderali: *RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens.* Bioinformatics, 25(5):678–9, 2009.

[118] Ripley, B.D.: *Modelling spatial patterns.* J.Royal Statistical Soc. Series B Stat. Methodol., 39:172–192, 1977.

[119] Ripley, Brian D.: *Spatial statistics.* John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[120] Rudnicka, D., J. Feldmann, F. Porrot, S. Wietgrefe, S. Guadagnini, M. C. Prevost, J. Estaquier, A. T. Haase, N. Sol-Foulon and O. Schwartz: *Simultaneous cell-to-cell transmission of human immunodeficiency virus to multiple targets through polysynapses.* Journal of virology, 83(12):6234–46, 2009.

[121] Scarselli, E., H. Ansuini, R. Cerino, R. M. Roccasecca, S. Acali, G. Filocamo, C. Traboni, A. Nicosia, R. Cortese and A. Vitelli:

The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus. EMBO J, 21(19):5017–25, 2002.

[122] Seidl, H., E. Richtig, H. Tilz, M. Stefan, U. Schmidbauer, M. Asslaber, K. Zatloukal, M. Herlyn and H. Schaider: *Profiles of chemokine receptors in melanocytic lesions: de novo expression of CXCR6 in melanoma.* Human pathology, 38(5):768–80, 2007.

[123] Sharma, S. D.: *Hepatitis C virus: molecular biology & current therapeutic options.* Indian J Med Res, 131:17–34, 2010.

[124] Sherer, N. M., M. J. Lehmann, L. F. Jimenez-Soto, C. Horensavitz, M. Pypaert and W. Mothes: *Retroviruses can establish filopodial bridges for efficient cell-to-cell transmission.* Nat Cell Biol, 9(3):310–5, 2007.

[125] Smolka, M. B., C. P. Albuquerque, S. H. Chen and H. Zhou: *Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases.* Proc Natl Acad Sci U S A, 104(25):10364–9, 2007.

[126] Snijder, B., R. Sacher, P. Ramo, E. M. Damm, P. Liberali and L. Pelkmans: *Population context determines cell-to-cell variability in endocytosis and virus infection.* Nature, 461(7263):520–3, 2009.

[127] Soriano, S. F., A. Serrano, P. Hernanz-Falcon, A. Martin de Ana, M. Monterrubio, C. Martinez, J. M. Rodriguez-Frade and M. Mellado: *Chemokines integrate JAK/STAT and G-protein pathways during chemotaxis and calcium flux responses.* European journal of immunology, 33(5):1328–33, 2003.

[128] Stark, C., B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski and M. Tyers: *The BioGRID Interaction Database: 2011 update.* Nucleic acids research, 39(Database issue):D698–704, 2011.

[129] Stelzl, U., U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn,

W. Birchmeier, H. Lehrach and E. E. Wanker: *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 122(6):957–68, 2005.

[130] Streib, K. and J.W. Davis: *Using Ripley's K-function to improve graph-based clustering techniques.* In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2305 –2312, june 2011.

[131] Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander and Jill P. Mesirov: *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 102(43):15545–15550, 2005.

[132] Supekova, Lubica, Frantisek Supek, Jongkook Lee, Shawn Chen, Nathanael Gray, John P. Pezacki, Achim Schlapbach and Peter G. Schultz: *Identification of human kinases involved in hepatitis C virus replication by small interference RNA library screening.* J Biol Chem, 283(1):29–36, Jan 2008.

[133] Suratanee, A., I. Rebhan, P. Matula, A. Kumar, L. Kaderali, K. Rohr, R. Bartenschlager, R. Eils and R. Konig: *Detecting host factors involved in virus infection by observing the clustering of infected cells in siRNA screening images.* Bioinformatics, 26(18):i653–8, 2010.

[134] Suzuki, Tetsuro: *A Hepatitis C virus-host interaction involved in viral replication: toward the identification of antiviral targets.* Jpn J Infect Dis, 63(5):307–311, Sep 2010.

[135] Szafran, Adam T., Maria Szwarc, Marco Marcelli and Michael A. Mancini: *Androgen receptor functional analyses by high throughput imaging: determination of ligand, cell cycle, and mutation-specific effects.* PLoS One, 3(11):e3605, 2008.

[136] Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen

and C. von Mering: *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 39(Database issue):D561–8, 2011.

[137] Tai, A. W., Y. Benita, L. F. Peng, S. S. Kim, N. Sakamoto, R. J. Xavier and R. T. Chung: *A functional genomic screen identifies cellular cofactors of hepatitis C virus replication.* Cell Host Microbe, 5(3):298–307, 2009.

[138] Tan, C. S.: *Sequence, structure, and network evolution of protein phosphorylation.* Science signaling, 4(182):mr6, 2011.

[139] Tan, C. S., B. Bodenmiller, A. Pasculescu, M. Jovanovic, M. O. Hengartner, C. Jorgensen, G. D. Bader, R. Aebersold, T. Pawson and R. Linding: *Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases.* Sci Signal, 2(81):ra39, 2009.

[140] Tan, C. S. and R. Linding: *Experimental and computational tools useful for (re)construction of dynamic kinase-substrate networks.* Proteomics, 9(23):5233–42, 2009.

[141] Timpe, J. M., Z. Stamataki, A. Jennings, K. Hu, M. J. Farquhar, H. J. Harris, A. Schwarz, I. Desombere, G. L. Roels, P. Balfe and J. A. McKeating: *Hepatitis C virus cell-cell transmission in hepatoma cells in the presence of neutralizing antibodies.* Hepatology, 47(1):17–24, 2008.

[142] Trotard, M., C. Lepere-Douard, M. Regeard, C. Piquet-Pellorce, D. Lavillette, F. L. Cosset, P. Gripon and J. Le Seyec: *Kinases required in hepatitis C virus entry and replication highlighted by small interference RNA screening.* FASEB J, 23(11):3780–9, 2009.

[143] Turner, M. and D. D. Billadeau: *VAV proteins as signal integrators for multi-subunit immune-recognition receptors.* Nature reviews. Immunology, 2(7):476–86, 2002.

[144] Tybulewicz, V. L.: *Vav-family proteins in T-cell signalling.* Current opinion in immunology, 17(3):267–74, 2005.

[145] Vaillancourt, F. H., L. Pilote, M. Cartier, J. Lippens, M. Liuzzi, R. C. Bethell, M. G. Cordingley and G. Kukolj: *Identification of a lipid kinase as a host factor involved in hepatitis C virus RNA replication.* Virology, 387(1):5–10, 2009.

[146] Vapnik, V.N.: *Statistical learning theory.* Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998.

[147] Veillette, A.: *Immune regulation by SLAM family receptors and SAP-related adaptors.* Nature reviews. Immunology, 6(1):56–66, 2006.

[148] Vila-Coro, A. J., J. M. Rodrguez-Frade, A. Martn De Ana, M. C. Moreno-Ortz, C. Martnez-A and M. Mellado: *The chemokine SDF-1alpha triggers CXCR4 receptor dimerization and activates the JAK/STAT pathway.* FASEB J, 13(13):1699–1710, 1999.

[149] Vinayagam, Arunachalam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E. Assmus, Miguel A. Andrade-Navarro and Erich E. Wanker: *A directed protein interaction network for investigating intracellular signal transduction.* Sci Signal, 4(189):rs8, Sep 2011.

[150] Vokes, M. S. and A. E. Carpenter: *Using CellProfiler for automatic identification and measurement of biological objects in images.* Curr Protoc Mol Biol, Chapter 14:Unit 14 17, 2008.

[151] Walter, T.; Held, M.; Neumann B.; Heriche J.-K.; Conrad C.; Pepperkok R. & Ellenberg J.: *A genome wide RNAi screen by time lapse microscopy in order to identify mitotic genes - computational aspects and challenges.* In *Proc. IEEE Internat. Symposium on Biomedical Imaging: From Nano to Macro (ISBI'2008)*, pages 328–331, 2008.

[152] Walter, Thomas, Michael Held, Beate Neumann, Jean-Karim Hrich, Christian Conrad, Rainer Pepperkok and Jan Ellenberg: *Automatic identification and clustering of chromosome phenotypes in a genome wide RNAi screen by time-lapse imaging.* J Struct Biol, 170(1):1–9, Apr 2010.

[153] Wang, Jun, Xiaobo Zhou, Pamela L. Bradley, Shih-Fu Chang, Norbert Perrimon and Stephen T C. Wong: *Cellular phenotype recognition for high-content RNA interference genome-wide screening.* J Biomol Screen, 13(1):29–39, Jan 2008.

[154] Wiles, Amy M., Dashnamoorthy Ravi, Selvaraj Bhavani and Alexander J R. Bishop: *An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme.* J Biomol Screen, 13(8):777–784, Sep 2008.

[155] Wixon, J. and D. Kell: *The Kyoto encyclopedia of genes and genomes–KEGG.* Yeast, 17(1):48–55, 2000.

[156] Wong, David W. S. and Jay Lee: *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS.* John Wiley & Sons, Hoboken, New Jersey, 2005.

[157] Wu, B., E. Y. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov and R. C. Stevens: *Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.* Science, 330(6007):1066–71, 2010.

[158] Xue, Y., Z. Liu, J. Cao, Q. Ma, X. Gao, Q. Wang, C. Jin, Y. Zhou, L. Wen and J. Ren: *GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection.* Protein Eng Des Sel, 24(3):255–60, 2011.

[159] Xue, Y., J. Ren, X. Gao, C. Jin, L. Wen and X. Yao: *GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy.* Molecular & cellular proteomics : MCP, 7(9):1598–608, 2008.

[160] Yang, Y.: *Wnts and wing: Wnt signaling in vertebrate limb development and musculoskeletal morphogenesis.* Birth defects research. Part C, Embryo today : reviews, 69(4):305–17, 2003.

[161] Yellaboina, S., A. Tasneem, D. V. Zaykin, B. Raghavachari and R. Jothi: *DOMINE: a comprehensive collection of known and predicted*

domain-domain interactions. Nucleic Acids Res, 39(Database issue):D730–5, 2010.

[162] Yoshimura, A., T. Naka and M. Kubo: *SOCS proteins, cytokine signalling and immune regulation.* Nature reviews. Immunology, 7(6):454–65, 2007.

[163] You, Xia, Annalee W. Nguyen, Abeer Jabaiah, Mark A. Sheff, Kurt S. Thorn and Patrick S. Daugherty: *Intracellular protein interaction mapping with FRET hybrids.* Proc Natl Acad Sci U S A, 103(49):18458–18463, Dec 2006.

[164] Zdobnov, E. M. and R. Apweiler: *InterProScan–an integration platform for the signature-recognition methods in InterPro.* Bioinformatics, 17(9):847–848, Sep 2001.

[165] Zernike, Fritz: *Beugungstheorie des Schneidenverfahrens und einer verbesserten Form, der Phasenkontrastmethode.* Physica, 1:689–704, 1934.

[166] Zhang, J., G. Randall, A. Higginbottom, P. Monk, C. M. Rice and J. A. McKeating: *CD81 is required for hepatitis C virus glycoprotein-mediated viral infection.* J Virol, 78(3):1448–55, 2004.

[167] Zhang, W., Y. Shao, D. Fang, J. Huang, M. S. Jeon and Y. C. Liu: *Negative regulation of T cell antigen receptor-mediated Crk-L-C3G signaling and cell adhesion by Cbl-b.* The Journal of biological chemistry, 278(26):23978–83, 2003.

[168] Zhang, Y., L. Tian, Y. Zheng, H. Qi, C. Guo, Q. Sun, E. Xu, D. Ma and Y. Wang: *C-terminal peptides of chemokine-like factor 1 signal through chemokine receptor CCR4 to cross-desensitize the CXCR4.* Biochemical and biophysical research communications, 409(2):356–61, 2011.

[169] Zhu, H., M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein and M. Snyder: *Global analysis of protein activities using proteome chips.* Science, 293(5537):2101–5, 2001.