From cancer gene expression to protein interaction: Interaction prediction, network reasoning and applications in pancreatic cancer

Dissertation

to receive the academic degree / zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

submitted to / vorgelegt an der
Dresden University of Technology
Department of Computer Science /
Technische Universität Dresden
Fakultät Informatik

by /eingereicht von

Dipl.-Inf. Gihan Elsir Ahmed Daw Elbait
born / geboren am 20. November 1975 in Omdurman/Sudan

defended on / verteidigt am

Dresden, 16. June 2009

Reviewers/Gutachter:
Prof. Dr. Michael Schroeder, Technische Universität Dresden, Inst. f. künstliche Intelligenz
Prof. Dr. Pedro Barahona, Universidade Nova de Lisboa, Departamento de Informática

To my beloved mother and father.

Abstract

Microarray technologies enable scientists to identify co-expressed genes at large scale. However, the gene expression analysis does not show functional relationships between co-expressed genes. There is a demand for effective approaches to analyse gene expression data to enable biological discoveries that can lead to identification of markers or therapeutic targets of many diseases. In cancer research, a number of gene expression screens have been carried out to identify genes differentially expressed in cancerous tissue such as Pancreatic Ductal Adenocarcinoma (PDAC). PDAC carries very poor prognosis, it eludes early detection and is characterised by its aggressiveness and resistance to currently available therapies. To identify molecular markers and suitable targets, there exist a research effort that maps differentially expressed genes to protein interactions to gain an understanding at systems level. Such interaction networks have a complex interconnected structure, whose the understanding of which is not a trivial task. Several formal approaches use simulation to support the investigation of such networks. These approaches suffer from the missing knowledge concerning biological systems. Reasoning in the other hand has the advantage of dealing with incomplete and partial information of the network knowledge.

The initial approach adopted was to provide an algorithm that utilises a network-centric approach to pancreatic cancer, by re-constructing networks from known interactions and predicting novel protein interactions from structural templates. This method was applied to a data set of co-expressed PDAC genes. To this end, structural domains for the gene products are identified by using threading which is a 3D structure prediction technique. Next, the Protein Structure Interaction Database (SCOPPI), a database that classifies and annotates domain interactions derived from all known protein structures, is used to find templates of structurally interacting domains. Moreover, a network of related biological pathways for the PDAC data was constructed.

In order to reason over molecular networks that are affected by dysregulation of gene expression, BioRevise was implemented. It is a belief revision system where the inhibition behaviour of reactions is modelled using extended logic programming. The system computes a minimal set of enzymes whose malfunction explains the abnormal expression levels of observed metabolites or enzymes.

As a result of this research, two complementary approaches for the analysis of pancreatic cancer gene expression data are presented. Using the first approach, the pathways found to be largely affected in pancreatic cancer are signal transduction, actin cytoskeleton regulation, cell growth and cell communication. The analysis indicates that the alteration of the calcium pathway plays an important role in pancreas specific tumorigenesis. Furthermore, the structural prediction method reveals ~ 700 potential protein-protein interactions from the PDAC microarray data, among them, 81 novel interactions such as: serine/threonine kinase CDC2L1 interacting with cyclin-dependent kinase inhibitor CDKN3 and the tissue factor pathway inhibitor 2 (TFPI2) interacting with the transmembrane protease serine 4 (TMPRSS4). These resulting genes were further investigated and some were found to be potential therapeutic markers for PDAC. Since TMPRSS4 is involved in metastasis formation, it is hypothesised that the upregulation of TMPRSS4 and the downregulation of its predicted inhibitor TFPI2 plays an important role in this process. The predicted protein-protein network inspired the analysis of the data from two other perspectives. The resulting protein-protein interaction network highlighted the importance of the co-expression of KLK6 and KLK10 as prognostic factors for survival in PDAC as well as the construction of a PDAC specific apoptosis pathway to study different effects of multiple gene silencing in order to reactivate apoptosis in PDAC.

Using the second approach, the behaviour of biological interaction networks using computational logic formalism was modelled, reasoning over the networks is enabled and the abnormal behaviour of its components is explained. The usability of the BioRevise system is demonstrated through two examples, a metabolic disorder disease and a deficiency in a pancreatic cancer associated pathway. The system successfully identified the inhibition of the enzyme glucose-6-phosphatase as responsible for the Glycogen storage disease type I, which according to literature is known to be the main reason for this disease. Furthermore, BioRevise was used to model reaction inhibition in the Glycolysis pathway which is known to be affected by Pancreatic cancer.

Published papers

1. Structural templates predict novel protein interactions and targets from pancreas tumour gene expression data.
   Gihan Dawelbait, Christof Winter, Yanju Zhang, Chrisitian Pilarky, Robert Grützmann, Jürg-Chrisitan Heinrich, and Michael Schroeder.
   Bioinformatics, 2007, Impact factor 2007: 5.0

2. Structural protein interactions predict kinase-inhibitor interactions in upregulated pancreas tumour genes expression data.
   Gihan Dawelbait, Christian Pilarsky, Yanju Zhang, Robert Grützmann, and Michael Schroeder.
   In Kay Diederichs, Robert Glen, and Oliver Kohlbacher, editors, Proceedings of the 1st International Symposium on Computational Life Science. Springer LNBI, 2005

3. Co-expression of KLK6 and KLK10 as prognostic factor for survival in pancreatic ductal adenocarcinoma.
   Felix Rückert, Mario Hennig, Constantina D. Petraki, Diana Wehrum, Marius Distler, Axel Denz, Michael Schroeder, **Gihan Dawelbait**, Holger Kalthoff, Hans-Detlev Saeger, Eleftherios P. Diamandis, Christian Pilarsky, and Robert Grützmann.
   British Journal of Cancer, 2009, Impact factor 2007: 4.6.

4. Conserved hydrogen bond patterns reveal structural and functional motifs in transmembrane protein regions.
   Annalisa Marsico, Andreas Henschel, **Gihan Dawelbait**, Christof Winter, Anne Tuukkanen, and Michael Schroeder.
   In Proceedings of Automated Function Prediction 2007.

5. Prova: Rule-based Java Scripting for Distributed Web Applications: A Case Study in Bioinformatics.
   Alex Kozlenkov, Rafael Penaloza, Vivek, Nigam, Loic Royer, **Gihan Dawelbait**, and Michael Schroeder.
   In Sebastian Schaffert, editor, Proceedings of Workshop on Reactivity on the Web at the International Conference on Extending Database Technology (EDBT 2006). Springer, 2006

6. Simultaneous gene silencing is an effective method for the reduction of apoptosis-resistance in pancreatic cancer cells.
   Felix Rückert, Nicole Samm, **Gihan Dawelbait**, Christof Winter, Arndt Hartmann, Michael Schroeder, Anne-Kathrin Lehner, Hans K Schackert, Ole Ammerpohl, Holger Kalthoff, Hans-Detlev Saeger, Robert Grützmann and Christian Pilarsky
   Submitted to Molecular Cancer

Paper 1 describes the algorithm and results of the protein-protein interaction predictions using a structural templates approach. This paper comprises the main scope of chapter 4. My contribution to the work in paper 1, is the construction and implementation of the algorithm along with the detailed investigation of the examples.

Paper 2 is an earlier version of the algorithm of paper 1 and only the results section of this paper is discussed in chapter four in the form of the kinases example.

Paper 3 is discussed in chapter four as case study III, my contribution was using the algorithm to predict interaction partners of KLK10 and KLK6. Paper 6 is discussed in chapter four as case study IV, my contribution was using the algorithm to predict interaction partners of apoptosis genes to aid the construction of a cancer related apoptosis pathway map.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The true impact of the genomics revolution is the transformation from an exclusively laboratory science into an information-rich science. The knowledge of full genomes made it possible to study patterns of gene expression under various conditions, by using tools like microarrays. The advent of microarray technology has created the possibility of monitoring the expression levels of thousands of genes in parallel. A common challenge faced by researchers is making sense out of such lists of differentially regulated genes for a better understanding of the underlying biological phenomena. An initial step towards this goal is the translation of the list of differentially expressed genes into a functional profile able to offer insight into the cellular mechanisms relevant in the given condition. However, such gene expression analysis does not show functional relationships between the elements of the co-expressed genes.

There is a demand for effective approaches to analyse gene expression data to enable biological discoveries that can lead to identification of markers or therapeutic targets of many lethal diseases. A variety of statistical and data mining techniques have been applied for the analyses of gene expression data. It is now understood that the combination of microarray data with other genomic or proteomic data is often able to provide a more comprehensive view of a biological system and facilitates an effective exploration of the data [Zhang, 2006]. Most crucial molecular bases of cellular operation such as metabolism, signalling, and regulation are largely sustained by different types of protein-protein interactions.

Another challenge in analysing large-scale expression data has been to extract biologically meaningful inferences from processes often represented as networks [Djebbari and Quackenbush, 2008]. There is a need to develop tools to understand not only the structure of the networks that exist, but also the rules that govern their behaviour and the interactions between the elements [Quackenbush, 2007].

In [Hanahan and Weinberg, 2000], the authors presented a cell map with the hallmark of cancer that covers the main knowledge of the current cancer interactome integrated into the network of biological pathways that are relevant to cancer diseases see Figure 1.1.

The aim of this work is to complement such efforts, by narrowing down the gap between the actual protein interactome and the incompleteness of the current sparse information about protein interactions due to the complexity and time consuming experimental techniques on that end. Mapping disease gene expression data into interaction and pathway networks for the in-

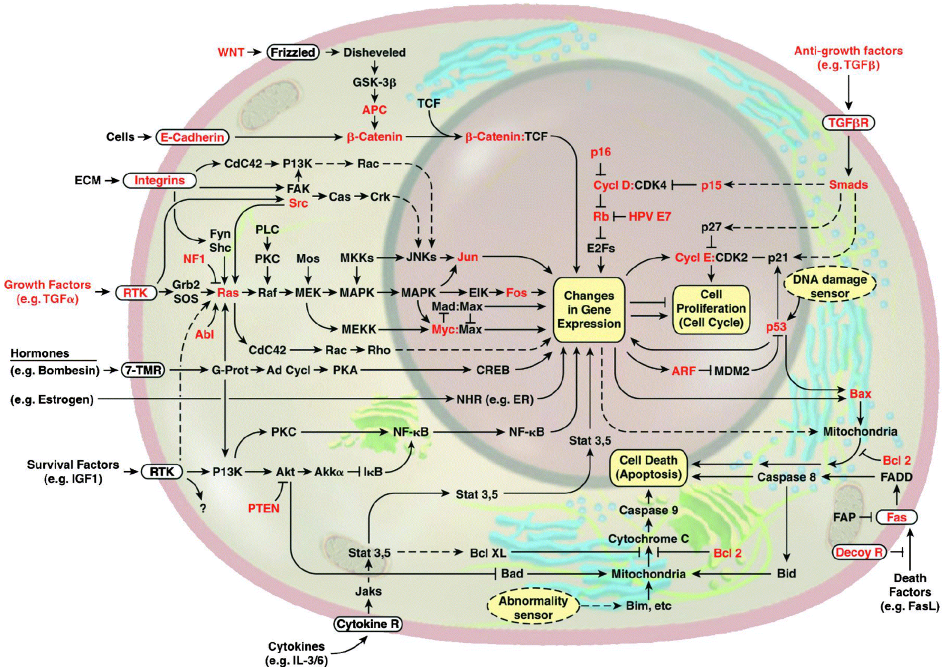


Figure 1.1: The Hallmarks of cancer from Hanahan and Weinberg [2000].

terpretation of such data can help us to reach functional understanding of ongoing processes of complex diseases from such networks. Furthermore, we need to consider the genes as switches whose dysregulation doesn't only affect their direct interacting partners but also the whole network of pathways they are involved in. Figure 1.2 illustrates the idea of this approach.

Modelling the behaviour of such interaction networks is a challenging yet a very important task. The use of logic programing to reason over such networks is a promising approach. Logical models are highly abstract, only a small amount of data is required for the modelling, they have a high analysis speed and the ability to perform inference in comparison to other standard modelling techniques such as continuous models [Karlebach and Shamir, 2008]. Belief revision has been widely used to model networks such as electric circuits [Damásio et al., 1997]. In the same manner, rules that governs biological interaction networks can be modelled so we can reason over them. The goal we hope to achieve from this large scale analysis is to provide a method that results in the development of new testable hypotheses of potential candidates as therapeutic and marker genes, that can be verified experimentally. In this chapter the open questions will be defined along with a brief summary of how they will be addressed.

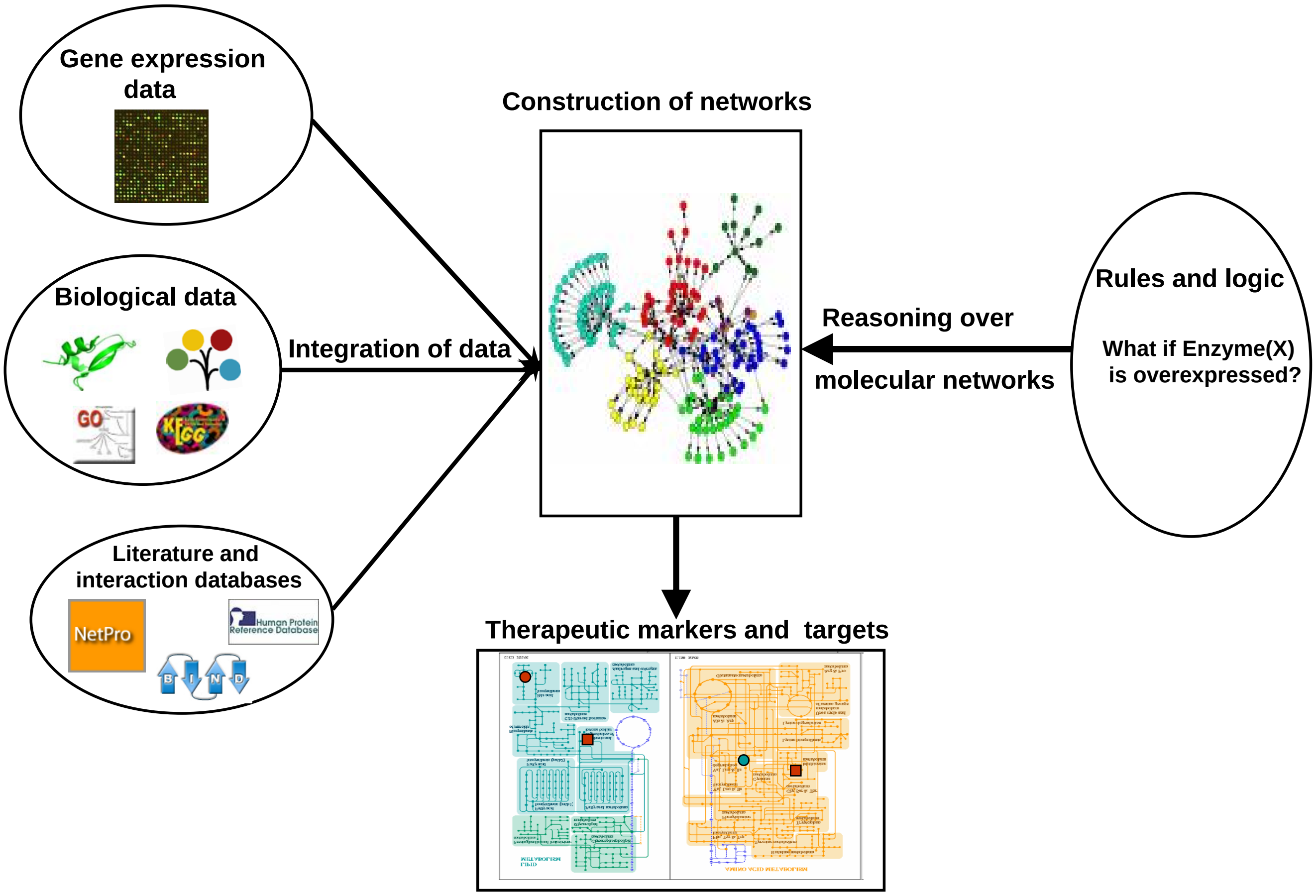


Figure 1.2: Approach: Complementing gene expression data with protein-protein interactions, biological pathways and literature extracted data for the construction of protein interactions networks to provide a more comprehensive view of a biological system and facilitates effective exploration of the data. Reasoning over these networks using rules that govern their behaviour and the interactions between the elements leads to the discovery of potential candidates as therapeutic and marker genes that should be verified experimentally.

1.1 Open research questions

1.1.1 Open problem 1: Can the use of protein-protein interaction data enhance knowledge discovery from gene expression data analysis to reveal potential markers and therapeutic targets for Pancreatic cancer?

Analysing high-throughput data produced by microarrays techniques is a challenging task. Gene expression profiling leads only to sets of potential candidate genes for further investigations, but it does not show functional relationships between co-expressed genes. One of the emerging principles in biology is that in most cases it is not the individual genes, but rather biological networks and pathways that derive the organism's response to a wide range of stimuli [Quackenbush, 2007].

Although closely related, gene expression and protein interaction data convey different biological meanings, and a coincident of interacting proteins and co-expressed genes is biologically significant [Zhang, 2006].

For constructing comprehensive interaction networks that reflects the hallmarks of complex diseases, it is necessary to integrate a variety of relevant annotation resources such as gene expression data, protein interaction, biological pathways and functional annotation of genes. It is argued that such networks can assist us to find potential markers, therapeutic targets and signature genes for cancer. There are several challenges that need to be tackled before the above mentioned question can be fully addressed:

Constructing interaction networks using structural interaction templates: Most proteins perform their functions after forming specific 3D structures. Thus, the structural identification of proteins is a crucial step towards determining their functions. Due to the difficult, time consuming and expensive standard experiments (X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy) used for structure recognition, it is still substantially lagging way behind the output of protein sequence data. In most high-throughput experiments there is usually little known about the genes produced. Therefore, there is a need for high-throughput prediction of protein structures. Among all the currently used computational methods, threading shows the most promise in the identification of structures. This is due to the fact that they are more sensitive than sequence alignment and can assign folds correctly even with low sequence similarity.

The interactions between proteins are important for many biological functions. Experimentally, interactions between pairs of proteins are inferred from yeast two-hybrid systems, affinity purification/mass spectrometry assays, or from protein microarrays. Current efforts are to develop computational methods for the prediction of protein interactions. Such methods have been developed to test whether interactions between homologous proteins can be modelled on the basis of an interaction of known structures.

Complementary annotation of the interaction network to assist knowledge discovery: One approach to extract meaning from lists of genes is to use annotation in a sophisticated way. To produce a more complete understanding to the biology behind the gene expression data we need to combine the interaction networks elements of the gene expression data to diverse sources of available information such as biological pathways, Gene Ontology annotations, and literature confirmed interactions.

Idea: As part of the current research project, we propose a method that uses inference from known structures, where structurally interacting domains act as templates for the predicted domains of the genes provided by gene expression data. The novelty of this method is that the conservation of the interface residues is considered as a quality control of our interactions. Furthermore, we will use the localisation of the proteins in the cell as a filter for interactions that are not likely to happen if found to be at different location in the cell. Connecting the genes through the construction of interaction networks facilitates the formulation of a more comprehensive picture of how the interacting genes influence each other. The aim is to show that the poor prognosis of lethal diseases such as pancreatic cancer can be improved by pointing out markers and signature genes, through the use of this type of interaction network approach.

1.1.2 Open problem 2: Does reasoning over molecular networks facilitate the analysis of gene expression data?

Due to the huge number of genes produced by microarrays screens, the molecular networks where these genes are involved can get very complicated. The main challenge posed by the construction of such networks is automating the reasoning step. Ideally, expert knowledge and logical reasoning should be used in answering questions such as “What if a change in a gene expression in the network is observed? How would that affect respective metabolic networks? How is this observation explained logically?”.

In order to answer such questions, the structure of molecular networks as well as the rules that govern their behaviour were investigated. Modelling these highly complex networks enables researchers to easily draw conclusions to explain expression profiles and their effect in metabolic pathways. Identification of metabolic pathways associated with cancer may have advantages for cancer control. There are considerable efforts directed at modelling biological networks, such as protein interaction and metabolic pathways. This modelling utilises techniques such as differential equations, rule based models with algebraic syntax, Hybrid Petri Nets and Hybrid Concurrent Constraint.

Idea: In the current approach we are aiming to use rules and reasoning to model a high level representation of inhibition of enzyme-catalysed reactions, which reasons over the KEGG network to explain abnormal behaviour of genes and metabolites expression level.

1.2 Outline

The major contribution of this thesis lies in the use of protein-protein interactions data and reasoning over molecular networks for the analysis of gene expression data to address the above mentioned open research problems. The rest of the thesis is organised as follows:

In the next chapter, an overview of the relevant literature and state of the art of standard approaches, for analysing gene expression data, with the emphasis on protein interaction networks, is presented.

The development of the research conducted in this thesis has progressed through several stages. Initially, I developed an algorithm based on domain-domain interactions as structural templates for predicting potential interactions within dysregulated genes from cancerous tissues. The different steps of the algorithm and the validation methods for filtering out false positives interaction from the resulting network is presented in Chapter 3.

Chapter 4 demonstrates four case studies where our protein-protein interaction predictions approach was applied to analyse pancreatic cancer gene expression data. In case study I, a novel pancreatic cancer network of known and predicted protein-protein interactions was constructed. The result of this study inspired the conduction of the other three case studies where the data sets were analysed from different perspectives. In case study II and III, potential therapeutic targets to fight chemoresistence and prognostic markers for enhancing survival in pancreatic cancer patients were studied. In case study IV, genes that were influencing the apoptotic pathway of the cancer cell were investigated.

Constructing predicted protein-protein interactions from high throughput data such as gene expression, is the first step towards assisting researchers to have a better picture over their data which initially was a list of co-expressed genes. The second step is to find causal functional relations between the elements of these data sets. For this purpose we make use of rules and reasoning to analyse molecular networks that are related to the Pancreatic cancer dysregulated genes. The method and two examples are presented in Chapter five.

Summary and discussion are presented in Chapter six.

Chapter 2

Background

2.1 A systems approach to pancreatic cancer

Understanding complex diseases such as cancer has been always an active research field. Several approaches that has the goal of providing the scientists with a catalogue of all human protein-protein interactions to study protein deregulation has been applied by many groups. For identifying disease genes and networks, [Goh et al., 2007] constructed a human disease network (Diseasome) where two disease genes are connected if they are associated with the same disorder. In [Freudenberg and Propping, 2002], Freudenberg and Propping, presented a method for prediction of disease relevant human genes from the phenotypic appearance of a query disease. Similarly [Aerts et al., 2006] prioritise candidate genes underlying biological processes or diseases, based on their similarity to known genes involved in these phenomena. For a convenient integration of functional annotation and statistical analysis of cancer related genes, the authors of [Dönnes et al., 2004] developed the cancer associated protein database (CAP) which demonstrated the success of integrative analysis approaches of cancer data.

In [Pospisil et al., 2006], the authors presented a combined approach to data mining of textual and structured data to identify cancer-related targets. Pospisil's approach is particularly interesting because the authors took a first step towards a systems biology approach by incorporating into their analysis functional annotations from the Gene Ontology [Harris et al., 2004] and relevant protein interactions from Ingenuity's Pathways Analysis. Recent databases such as pSTIING [Ng et al., 2006], and Cyclonet [Kolpakov et al., 2007] focus on integrating and linking cancer gene expression data to pathways and interaction databases. [Rhodes et al., 2005] initiated this line of thinking by building a probabilistic network model, which is based among others on co-expression, and by identifying relevant interactions for pancreatic cancer. Other groups have hypothesised that a more effective means of marker identification may be to combine gene expression measurements over groups of genes that fall within common pathways and human phenome as discussed in [Chuang et al., 2007, Marc A van Driel et al., 2006].

Over the past years, such a network-based approach has become possible. Fuelled by high-throughput interaction experiments [Uetz et al., 2000, Ito et al., 2001, Rain et al., 2001, Gavin et al., 2002, Ho et al., 2002, Giot et al., 2003, Li et al., 2004, Rual et al., 2005, Gavin et al., 2006], large databases with thousands of interactions have emerged such as IntAct [Hermjakob et al., 2004], STRING [von Mering et al., 2007], DIP [Xenarios et al., 2000], HPRD [Peri et al., 2003], BIND [Bader and Hogue, 2000], KEGG [Ogata et al., 1999], and Reactome [Tope et al., 2005]. They have been complemented by databases for structural interactions such as PIBASE [Davis and Sali, 2005], PSIBASE [Gong et al., 2005], 3did [Stein et al., 2005], and SCOPPI [Winter et al., 2006b]. Finally, there are many efforts to extract interactions from literature, among them iHOP [Hoffmann and Valencia, 2004] and ALI BABA [Plake et al., 2006].

Here, we follow Rhodes et al. and Pospisil et al. taking a network centric approach to the reconstruction of signalling cascades and the identification of promising targets. We go beyond this work by including into our networks predicted interactions based on structural templates, which help elucidating the mode of interaction of deregulated proteins. Ultimately, the aim is to identify drug targets that explain the mechanism of action of existing and novel drugs.

This chapter is intended to provide definitions of the basic components of our integrative approach. The chapter also provides a review of the state of the art of relevant efforts in the fields of pancreatic cancer, gene expression analysis, protein-protein interactions, biological pathways and reasoning over networks.

2.2 Definitions

2.2.1 Pancreas cancer

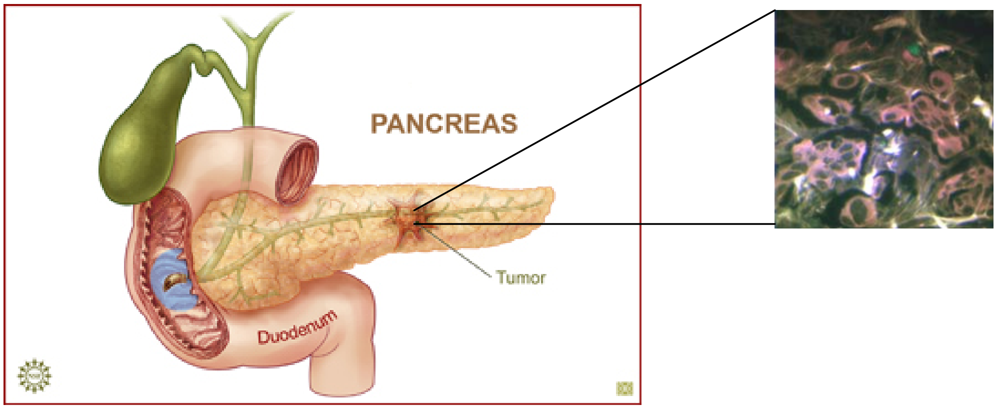


Figure 2.1: A sketch showing the pancreas location in the human body and a pancreatic cancer tissue with the cancerous cells in dark pink

Initially, I will start by defining the pancreas organ, the location from which pancreatic cancer originates. The pancreas is a small organ located in the upper abdomen in close proximity to the duodenum. It serves two major functions: The pancreatic exocrine cells (takes up the vast majority of the tissue mass of pancreas) produce digestion enzymes also known as the digestive juices that are secreted in response to food intake. The endocrine cells of the pancreas are scattered throughout the organ in groups called the islets of Langerhans. These cells secrete Insulin, Glucagon and Somatostatin. The two main diseases of the pancreas are pancreatitis and pancreatic cancer. Pancreatic tumours are classified as either exocrine or endocrine depending on which type of tissue they arise from within the gland. Pancreatic cancer is the fourth leading cause of death due to cancer in virtually all industrialised countries, and causes more than 34,000 deaths per year in the United States[1]. It is most common in blacks, in men, and in patients with either diabetes or heredity chronic pancreatitis. Pancreatic cancer is difficult to detect, hard to diagnose, early to metastasise, and resistant to treatment. These four characteristics of pancreatic cancer and their synergistic interactions contribute to the high mortality and short life expectancy making it one of the deadliest of all cancers. Pancreatic ductal adenocarcinome (PDAC) is the most common pancreatic neoplasms and accounts for between 80% and 90% cases of pancreatic tumour [Hezel et al., 2006a], it has an extremely poor prognosis. To improve the prognosis, novel molecular markers and targets for earlier diagnosis and adjuvant and neoadjuvant treatment need to be identified. Despite the progress made in recent years in the treatment of various types of cancer, the dismal prognosis of PDAC remains unchanged. Apart from surgery there is no curative therapy, and even resected patients usually die within 1 year of the operation. In this situation there is an urgent need to understand more about the causes and the pathogenesis of PDAC [Dawelbait et al., 2005].

2.2.2 Microarray and gene expression data

A fundamental step in analysing any complex diseases is to identify the genes associated with this disease. This can be performed by determining a list of co-expressed genes and their differentially expressed levels from a diseased tissue.

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, where single strands of complementary DNA for the genes of interest are placed on spots arranged in a grid, typically a glass slide. From a sample of interest, e.g. a tumour biopsy, the mRNA is extracted, labelled and hybridised to the array. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample, see Figure 2.2. The result of such experiments are usually in the form of a matrix of gene expression levels. This matrix describes three possible states, red for overexpressed, green for underexpressed and black for no change in expression as shown in Figure 2.3 (b). These methods are applied by several groups to identify general and specific PDAC genes from gene expression profiles obtained from pancreatic cancers to determine those genes most differentially expressed and thus with the most promise for translation into clinically useful targets. Gene expression profiling using high-throughput microarray experiments result in huge amounts of data, which still need to be interpreted. Due to the biological complexity of gene expression, it is very hard to reproduce the same experimental results. This is due to several factors such as histology, number of samples, microdisection, and the use of different array technology [Grutzmann et al., 2004b]. Therefore,

[1] http://www.pancreas.org

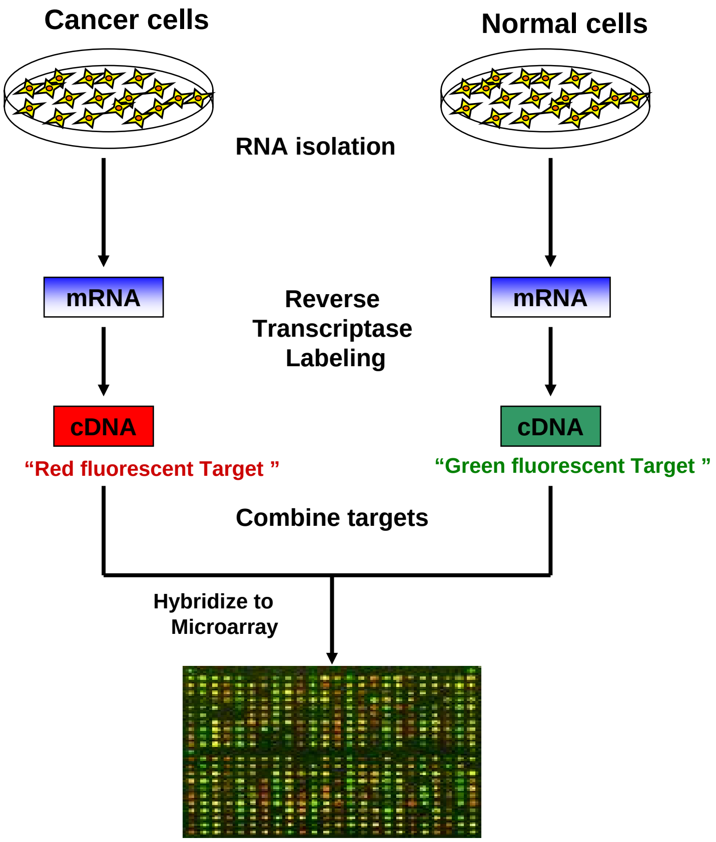


Figure 2.2: Microarray schema

biologists conducting such experiments should stick to standard experimental design and to the ”Minimum Information About a Microarray Experiment” (MIAME) checklist [Brazma et al., 2001]. The lack of standardisation in arrays presents an interoperability problem in bioinformatics, which hinders the exchange of array data. The quality of the data produced is of critical importance if statistically and biologically valid conclusions are to be drawn from the data. The analysis of DNA microarrays poses a large number of statistical problems, including the normalisation of the data.

The standard approaches used to analyse such high-throughput data are to cluster its elements into functional categories using different types of gene annotation, for example gene annotation or pathway information from biological pathway databases (i.e KEGG see Figure 2.3 (a)).

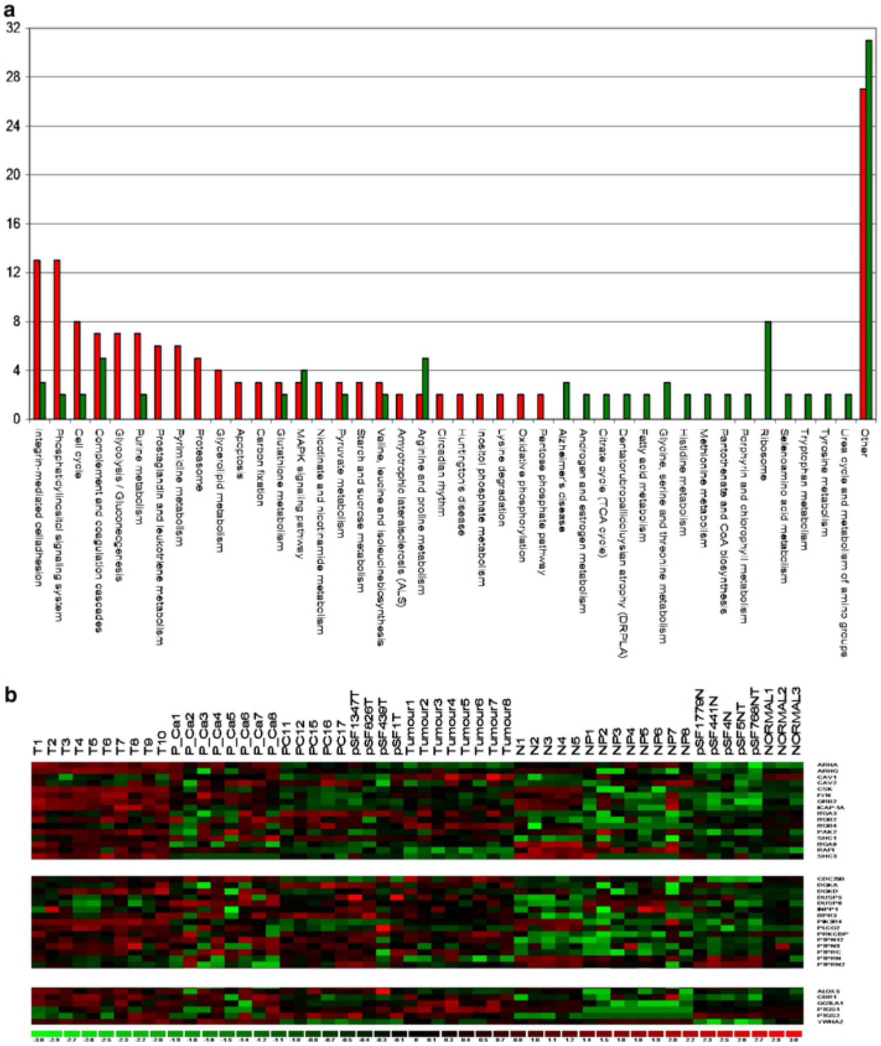


Figure 2.3: (a) Distribution of the annotated genes into KEGG functional categories. (b) Individual expression levels of the genes from [Grutzmann et al., 2005]. Red: genes overexpressed in PDAC; green: genes underexpressed in PDAC.

2.2.3 Prediction of Protein-Protein interactions

The clustering techniques applied on the huge sets of gene expression data could only provide correlation between these genes. But, in order to find mechanism of action of these genes when interacting together, we need to find causal relationships that link these genes to cause such complex diseases. This could only be provided by the completion of the interactome. At the moment, the current knowledge of interacting proteins is sparse and only methods of prediction of such interactions can help us to construct a more detailed networks that illustrates the hallmarks of complex diseases.

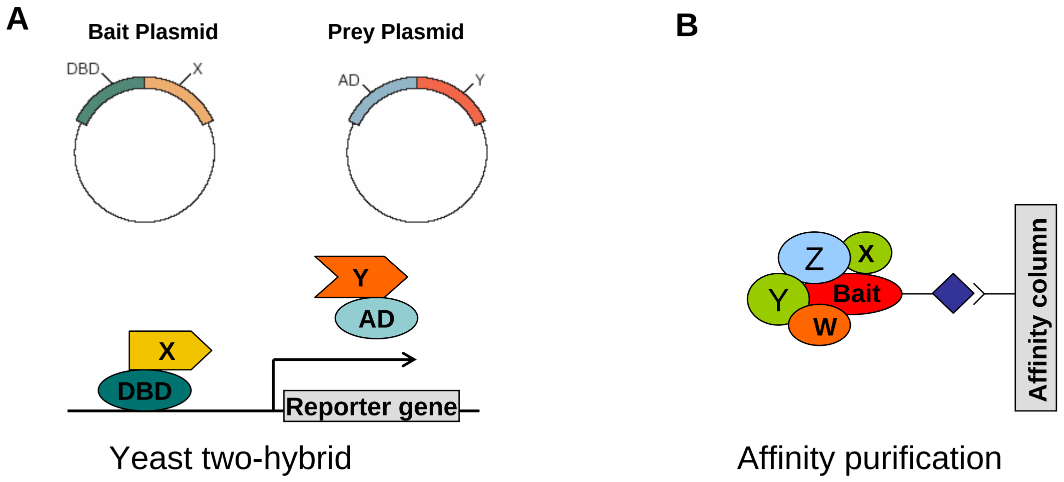


Figure 2.4: A sketch showing the (A) Yeast two-hybrid system and (B) Affinity purification schema

Protein-protein interactions are essential in almost all biological processes, extending from the formation of cellular macromolecular structures and enzymatic complexes to the regulation of signal transduction pathways. Interacting proteins are more likely to be involved in similar biological functions and processes and thus they are more likely to be co-expressed. Thus, one central task in the study of a protein is to determine its interaction partners. The most widely used experimental methods to determine protein-protein interactions are

Experimental Methods :

Yeast two-hybrid: The yeast two-hybrid [Fields and Song, 1989] system is based on the fact that eukaryotic transcriptional factors consist of two individual domains the DNA-binding domain(DBD) and its activation domain (AD). Each of these two parts is fused to the proteins of interest (X and Y), only if proteins X and Y interact with one another are the DBD and AD brought together to activate the expression of the reporter gene as shown in Figure 2.4 (A).

Affinity purifications: In Affinity purification [Puig et al., 2001] the protein of interest, known as the bait,is displayed as dark purple in Figure reffig:exp (B). The tagged protein is then purified with its interacting partners(W-Z), usually identified by mass spectrometry.

X-ray crystallography and NMR: provide an atomic description of the binding sites of interacting proteins. X-ray crystallography provides atomic resolution models for protein and complexes. NMR on the other hand defines interaction interfaces between proteins for which 3D structures are known [Aloy and Russell, 2006].

Computational Methods:
In the post genomic era, the importance of protein-protein interaction is becoming even more apparent. The existing drafts of organism interactomes from high throughput protein interaction approaches are still far from complete and need therefore to be complemented by computational predictions. For this purpose a variety of computational approaches has been developed [Valencia and Pazos, 2002].

They can be grouped into sequence based and structure based methods. The former include phylogenetic profiling [Pellegrini et al., 1999, Sun et al., 2005], genomic context analysis [Galperin and Koonin, 2000], gene fusion [Marcotte et al., 1999], sequence signatures [Sprinzak and Margalit, 2001], linear motifs [Puntervoll et al., 2003, Neduva and Russell, 2006]. Several studies made use of homologous interactions in other species to predict protein interactions [Ben-Hur and Noble, 2005, Espadaler et al., 2005, Kim et al., 2004, Han et al., 2004]. Table 2.1 provides brief definitions of the mentioned computational protein-protein interaction predictions methods. Machine learning approaches fall into both categories, depending on whether the describing features are derived from known structures or sequences. Structure based methods are briefly described below.

Structure based Machine Learning techniques: Given interacting and non-interacting surface patch data, a mapping into feature space is defined. The described features include surface conservation, curvedness, hydrophobicity, shape complementarity and physicochemical surface properties. Support vector machines are frequently applied [Bradford and Westhead, 2005, Bordner and Abagyan, 2005, Koike and Takagi, 2004].

Inference from known structures: is based on the assumption that similar sequences have similar folds and that domains with a similar fold interact through the same surface [Aloy and Russell, 2006]. The principle is illustrated in Figure 2.5.

Structural templates: A database of structural interface templates of interacting proteins has been compiled by [Ogmen et al., 2005]. Query proteins can then be compared against these templates Aytuna et al. [2005]. Two query templates are inferred to be interacting, if they can be respectively aligned to a pair of interface templates.

Although sequence based method were successfully used to predict protein interactions, most of them require the knowledge of whole genomes. In the other hand, when studying or modelling biological systems, a full understanding of how molecules interact comes only from three-dimensional (3D) structures, as they provide crucial atomic details about binding [Aloy and Russell, 2006]. In our approach we utilise a structural based interaction prediction method, that goes beyond inference from known structures by considering conservation of interacting interfaces.

2.2.4 Biological pathways

to understand biological processes we equally require the knowledge of direct and indirect relationships, that can include interaction partners, shared pathway, same cellular localisation or similar tissue specific expression levels.

	Prediction Method	Reference	Definition	Consider full genomes	Consider structure	Interface conservation	Homology modelling
	Phylogenetic profiling	Pellegrini et al. [1999], Sun et al. [2005]	Identify the presence or absence of patterns of domains in species	✓	×	×	×
	Genomic context analysis	Galperin and Koonin [2000]	Deals with frequently reoccurring neighbourhood relations of pairs of proteins on chromosomes in different species	✓	×	×	×
Sequence based	Sequence signatures	Sprinzak and Margalit [2001]	Predict interactions using inference from domain-pairs that contain sequence signatures that are known to be interaction-mediating in other proteins	✓	×	×	×
	Gene fusion	Marcotte et al. [1999]	It exists between genes that often interacted together and thus were conveniently combined to a single gene with multiple domains	✓	×	×	×
	Linear motifs	Puntervoll et al. [2003], Neduva and Russell [2006]	Are interfaces of short stretches of loop regions that are sequentially consecutive and able to arrange its structure	✓	×	×	×
	Machine Learning techniques	Bradford and Westhead [2005], Bordner and Abagyan [2005], Koike and Takagi [2004]	Extracte Features from protein surface patches. A positive data sample for a protein-protein interaction is a patch or segment that is involved in the interaction	×	✓	×	✓
Structure based	Inference from known structures	Aloy and Russell [2006]	Is based on the assumption that domains with similar folds interact through the same surface	×	✓	×	✓
	Structural templates	Aytuna et al. [2005]	Two query templates are inferred to be interacting, if they can be respectively aligned to a pair of interface templates	×	✓	×	✓

Table 2.1: Sequence and structure based protein protein prediction methods

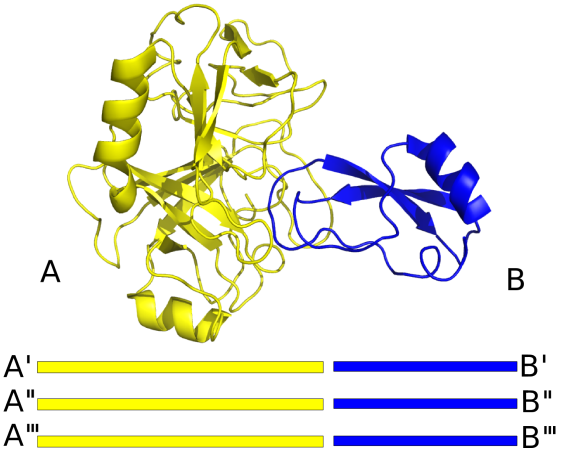


Figure 2.5: Interaction inference from known structures: a structure for the interaction of a protease (domain A, shown in yellow) and a Kunitz-type inhibitor (domain B, shown in blue) exists (PDB: 1brc). Other protease — Kunitz-type-inhibitor sequence pairs ($A' - B'$, $A'' - B''$, $A''' - B'''$) can be assumed to interact as well.

Biological pathways represent networks of complex reactions within a cell catalysed by enzymes, resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway (then called a *flux generating step*). Common properties of metabolic pathways includes reversibility of reactions, regulation of pathways using cycles or feedback inhibition in living cells. Biological pathways also model how biological molecules interact to accomplish a biological function and to respond to environmental stimuli [Saraiya et al., 2005]. Pathways capture the current knowledge of biological processes and are derived through scientific experimentation and data analysis. They are usually used to summarise the results of thousands of experiments in order to describe the flow of signals and metabolites in the cell. Several databases of metabolic and signalling pathways were produced during the past decade, such as BioCyc [Karp et al., 2005], BioCarta[2], Reactome [Joshi-Tope et al., 2005], BRENDA[3] which focuses on enzymatic catalysis, MetaCyc [Caspi et al., 2006] on metabolic pathways, aMAZE[4], a relational database for pathways and cellular process, The Enzyme and Metabolic pathways database EMP[5] focuses as well on enzymes and metabolic pathways, PathDB [Blanchard et al., 2000] has a wider span of interest which includes enzymes, pathways, kinetic, thermodynamic properties of pathway components, and finally Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al., 2002] which contains data about genes, enzyme, metabolic/signaling/regulatory reactions and visual maps with coordinates. In general, these resources represent the relationships between molecules in a cell either

[2] http://www.biocarta.com

[3] http://www.brenda.uni-koeln.de

[4] http://www.amaze.ulb.ac.be

[5] www.biobase.com/EMP

Pathway database	Reference	Features
MetaCyc	Caspi et al. [2006]	Metabolic pathways and enzymes (curated from literature).
BioCyc	Karp et al. [2005]	Uses MetaCyc to predict metabolic networks of 350 organisms. It includes metabolic pathways, enzymes, metabolites and reactions.
BioCarta	http://www.biocarta.com	Biological pathways.
Reactome	Joshi-Tope et al. [2005]	Biological pathways.
BRENDA	http://www.brenda.uni-koeln.de	Enzymatic catalysis.
aMAZE	http://www.amaze.ulb.ac.be	Pathways and cellular process.
EMP	www.biobase.com/EMP	Enzymes and metabolic pathways.
PathDB	Blanchard et al. [2000]	Enzymes, pathways, kinetic, thermodynamic properties of pathway components.
KEGG	Kanehisa et al. [2002]	Genes, enzyme, metabolic/signaling/regulatory reactions and visual maps with coordinates.

Table 2.2: Databases of metabolic and signalling pathways

as reactions or as activation or inhibition events [Aloy and Russell, 2006]. By understanding the dynamics of such pathways computational models that predict the mechanisms of diseases that occur when the cellular processes are dysregulated, can be developed. This will have significant impacts on biotech applications and drug discovery as this prediction can replace tedious and costly lab experiments [Karlebach and Shamir, 2008].

Various computational methods have been developed for the modelling and analysis of pathways. The most frequently used standard models can be divided into two classes:

Logical models: were introduced in the 70's by Kauffman and Thomas [Glass and Kauffman, 1973, Thomas, 1973]. It is a modelling method that is discrete and logic-based. Logical models provide a basic understanding of the different functionalities of a given network under varying conditions. They are highly abstract and hence require the least amount of data for modelling. Their qualitative nature makes them flexible and easy to fit to biological phenomenon. These models can be analysed using a number of well established mathematical methods such as Boolean networks [Kauffman, 1969] and Petri nets [Petri, 1962]. Logical models requires discretisation of the real valued data, at the cost of reducing the accuracy of the data.

Continuous models: incorporate real-valued parameters produced by biological experiments (e.g reaction rates, cell mass and gene expression intensities) over a continuous timescale. They allow a straightforward comparison of the global state and experimental data. However, these parameters of the continuous models are based on estimations since the quantitative measurements cover only a fraction of the system entities. Some continuous models make use of Ordinary Differential Equations (ODE) which was suggested by [Goodwin, 1963]. It provides detailed information about the network dynamics, but requires high quality data on kinetics parameters which makes it applicable only to a small number of systems.

2.2.5 Belief revision for rule based networks modelling

The task of incorporating a new fact into an existing knowledge base of a certain domain is called belief revision.

In particular, a belief revision occurs when a new piece of information that is *inconsistent* with the present belief system (or database) is added to that system in such a way that the result is a new consistent belief system. In a generic belief revision system the database is not viewed merely as a collection of logically independent facts, but rather as a collection of axioms from which other facts can be derived. It is the interaction between the updated facts and the derived facts that is the source of the problem.

2.3 Review of the state of the art

2.3.1 Recent advances in analysis of gene expression data

Gene expression profiling by means of DNA microarrays is a powerful technique to simultaneously screen thousands of genes in a cell population and is used to identify the mechanisms of deregulated molecular functions in pancreatic carcinoma cells. Many researchers have carried out gene expression experiments coupled to computational analyses to identify relevant markers and targets for pancreatic cancer. Iacobuzio-Donahue et al. [2002] used the Gene Logic Inc[6] BioExpress platform - a genomic database of gene expression data - for discovering novel tumour markers of pancreatic cancer. In a similar manner, Iacobuzio-Donahue et al. [2003b] performed a comprehensive evaluation and comparison of PDAC gene expression data where the authors used dimension reduction with Principal Components Analysis (PCA). Among the most differentially expressed genes identified by PCA were Mesothelin, Muc4, Muc5A/C, Kallikrein 10, Transglutaminase 2, Fascin, TMPRSS3 and stratifin. Likewise, Hustinx et al. [2004] used Serial Analysis of Gene Expression (SAGE) to identify PDAC genes. The differential expression of seven genes, involved in multiple cellular processes such as signal transduction (MIC-1), differentiation (DMBT1 and Neugrin), immune response (CD74), inflammation (CXCL2), cell cycle (CEB1) and enzymatic activity (Kallikrein 6), were experimentally validated.

Aguirre et al. [2004] used array comparative genomic hybridisation (CGH) on a cDNA microarray platform to define the copy number alterations (CNAs) in a panel of pancreatic adenocarcinoma cell lines and primary tumour specimens. A number of research groups are now developing methods that integrate interaction data with gene expression profiles and subcellular localisations and literature mining [Jansen et al., 2003, Mering et al., 2005].

[6] www.genelogic.com/

2.3.1.1 Analysis using the Gene ontology

Analysis of microarray data most often produces lists of genes with similar expression patterns, which are then subdivided into functional categories for biological interpretation. This is done by grouping co-expressed genes using their annotations. Such functional categorisation is most commonly accomplished using Gene Ontology (GO) categories. Ontologies can help in identifying and clustering sequence data that share common characteristics. GO [Ashburner et al., 2000] is widely used in biology to annotate sequence and structure data. It follows a human annotation process that is performed by the field experts by consulting relevant literature. It contains $\sim$ 30000 terms on biological process (BP), molecular functions(MF) and cellular component (CC). As a result of the GO initiative lots of annotation databases such as GOA [Camon et al., 2004] were established. Even though, GO has become the de facto standard and is used for the annotation of many biological databases, it has been criticised for numerous reasons. In [Smith et al., 2003, Kumar and Smith, 2004, Smith, 2004, Smith et al., 2004, Schulz et al., 2005, Kumar and Smith, 2003, Rosa and Smith, 2004, Kumar et al., 2004]. The authors discussed the problems of the existing design of GO emphasising on the point that GO does not have a formal architecture for its parts "is-a" and "part-of" relations, as well as how the terms of the three GO separate ontologies CC, BP, MF relate to each other. On the the other hand, [Bada et al., 2004] described GO as a success story, and presented an application which used GO to find relations between sequence similarity and semantic similarity [Lord et al., 2002, Hennig et al., 2003a,b]. Functional annotation of differentially expressed genes is a necessary and a critical step in the analysis of microarray data. Considerable effort was exerted to develop a large variety of tools for interpreting large lists of genes produced by high-throughput experiments (Micrroarrays and RNAi). A collection of tools that maps predominant functional themes of a given gene set on the GO hierarchy are listed on the GO website. A selection of these tools which is geared towards interpreting gene expression data is presented in the subsequent section. The Biological Networks Gene Ontology tool (BiNGO) [Maere et al., 2005]. NetAffx [Cheng et al., 2004] is a web based interactive tool that permits the traversal of the GO graph in the context of microarrays. It accepts as an input a list of Affymetrics sets and outputs a GO interactive graph in the form of a colour-map/heat-map, coloured according to the significance of the measurements. ChipInfo [Zhong1 et al., 2003] is designed for retrieving annotations from online databases (NetAffx and GO) and organising the information into an easily interpretable output format. GOAL, the GO Automated Lexicon [Volinia et al., 2004] is a web based application for the identification of functions and processes regulated in microarray and SAGE experiments. GO-Mapper [Smid and Dorssers, 2004] is a tool that quantitatively link gene expression levels to GO for multiple experiments in an automated way. Karma [Cheung et al., 2004] is a web server application for comparing and annotating heterogeneous microarray platforms. GOtcha [Martin et al., 2004] is a new method for the prediction of protein functions assessed by the annotation of seven genomes. GOTree Machine [Zhang et al., 2004] (GOTM) is a web-based platform for interpreting sets of interesting genes using GO hierarchies. GOSTAT [Beissbarth and Speed, 2004] is a tool that finds statistically overrepresented GO within a group of genes. GeneInfoViz [Zhou and Cu, 2004] is a tool for constructing and visualising gene relation networks. GoMiner [Zeeberg et al., 2003] is a resource for biological interpretation of genomic and proteomic data. GoMiner classifies the genes into biologically coherent categories and assesses these categories.

Go Tool	Reference	Features
BiNGO	Maere et al. [2005]	Determine which GO categories are statistically over-represented in a set of genes.
NetAffx	Cheng et al. [2004]	GO interactive graph in the form of a colour-map/heat-map.
ChipInfo	[Zhong1 et al., 2003]	Retrieve and organise annotations from (NetAffx and GO) into an easily interpretable output format.
GOAL	Volinia et al. [2004]	Identify of functions and processes regulated in microarray and SAGE experiments.
GO-Mapper	Smid and Dorssers [2004]	Analyse gene expression data using the expression level as a score to evaluate Gene Ontology terms.
Karma	Cheung et al. [2004]	Compare and annotate heterogeneous microarray platforms.
GOtcha	Martin et al. [2004]	Predict protein functions assessed by the annotation of seven genomes.
GOTree Machine	Zhang et al. [2004]	Interpret sets of genes using GO hierarchy.
GOSTAT	Beissbarth and Speed [2004]	Determine Identify statistically significant over- or under-represented GO categories within lists of genes.
GeneInfoViz	Zhou and Cu [2004]	Construct and visualise gene relation networks.
GoMiner	Zeeberg et al. [2003]	Organise lists of genes from a microarray experiment for biological interpretation in the context of the Gene Ontology.

Table 2.3: GeneOntolgy Tools for analysing gene expression data

A study that proposes that transitive expression similarity among genes can be used as an important attribute to link genes of the same biological pathway [Fages et al., 2005]. A complete list of tools that uses GO for analysis of data sets including gene expression and microarray data is listed in `http://www.geneontology.org/GO.tools.microarray.shtml`

2.3.1.2 Analysis using data mining

To be able to sort and analyse huge amounts of data, sophisticated data mining algorithms that identify trends within data, that go beyond simple analysis are needed. Data mining, databases and bioinformatics are frequently used to identify cancer-related targets. In [Pospisil et al., 2006] a new, rapid, data mining strategy that is based on a combination of curated knowledge bases such as (NCBI Genomic Biology, Ensembl and UCSC Genome Browser) with protein databases (UniProt - the universal protein resource) [Bairoch et al., 2004] and the RCSB Protein Data Bank (PDB , the database of protein structures) [Berman et al., 2000] and derived databases such as EMBL-EBI InterPro [Hunter et al., 2008] (database of protein families, domains and functional sites), has revealed the unique characteristics of several programs, including the entity retrieval capability and the Gene Ontology term-filtering specification of LS-Graph[7] and the full-text-based knowledge of Ingenuity Pathway Analysis[8]. The study revealed four interesting results, which are Alkaline phosphatase (various cancers), prostatic acid phosphatase, prostate-specific antigen (prostate cancer), and extracellular sulfatase 1 (pancreatic cancer). Cao et al. [Cao et al., 2004] applied another bioinformatics analyses to characterise ESTs that were found to be highly overexpressed in a series of pancreatic adenocarcinomas. The authors used basic local alignment search tools BLAST [Altschul et al., 1990], BLASTN[9], and BLASTX[10], for identifying protein coding genes corresponding to the ESTs. Subsequently, in order to pick the most relevant candidate genes for a more detailed analysis, the authors looked for domains/motifs in the open reading frames using SMART [Letunic et al., 2006] and Pfam [Finn et al., 2007]. The differential expression of a subset of genes was experimentally confirmed at the protein level by immunohistochemical labelling of tissue microarrays (inhibin beta A [INHBA] and CD29) and/or at the transcript level by RT-PCR (INHBA, AKAP12, ELK3, FOXQ1, EIF5A2, and EFNA5).

2.3.2 Literature confirmed cancer genes

Here, a number of previously reported differentially expressed cancer genes [Higgins et al., 2007] are listed, among them are the K-ras oncogene whose mutation has been identified in 90% of pancreatic cancers, the insulin-like growth factor (IGFBP4/5) and STAT1 a signal transducers and activators of transcription family.

SMADS are proteins of the TGFβ signalling pathway. Downregulation or loss of SMAD4 was shown to be important for pancreatic carcinogenesis. Sato et al. identified that epigenetic inactivation of Tissue factor pathway inhibitor 2 TFPI2 is a common mechanism that contributes to the aggressive phenotype of pancreatic ductal adenocarcinoma. Sova et al. identified TFPI2 as a biomarker that is repressed in cervical cancer. TMPRSS4 has been also identified as a biomarker for thyroid cancer [Kebebew et al., 2005]. Furthermore, [Mertz et al., 2007], identified recurrent gene fusions of TMPRSS2, a paralog of TMPRSS4, that mediate the overexpression of ETS transcription factor family members, most commonly ERG in prostate cancer. SERPINI2 a protease inhibitor is located at the chromosomal position 3q26.1-q26.2, a region that has been linked to a genetic risk for breast cancer. [Ozaki et al., 1998] has also shown that down-regulation of SERPINI2 may play a significant role in development or progression of pancreatic cancer. The increase of expression of CD44, a transmembrane protein involved in cell-to-matrix interactions, promotes metastatic potential of pancreatic carcinoma cells [Coppola, 2000]. The FOXM1 gene is upregulated in pancreatic cancer and basal cell carcinoma due to the transcriptional regulation by Sonic Hedgehog (SHH) pathway [Katoh and Katoh, 2004]. BRCA1, whose mutation appears to confer increased susceptibility for PDAC [Hezel et al., 2006b], as well as STK11, which is a tumour suppressor gene, was found to be involved in regulation of diverse processes such as cell polarity and metabolism.

[7] http://lsgraph.it-omics.com/

[8] http://www.pir.uniprot.org/

[9] http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn

[10] http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Translations&PROGRAM=blastx

Some of the above identified genes were investigated as therapeutic targets. Fleming et al. provided support that silencing mutant K-ras through RNA interference results in alteration of tumour cell behaviour in vitro and suggests that targeting mutant K-ras specifically might be effective against pancreatic cancer in vivo. [Lebedeva et al., 2006] as well targeted K-ras by using an adenovirus expressing a novel cancer-specific apoptosis-inducing cytokine gene.

Taniuchi et al. identified KIF20A as a candidate for development of drugs to treat PDACs. Knockdown of endogenous KIF20A expression in PDAC cell lines by small interfering RNA drastically attenuated growth of those cells, suggesting an essential role for the gene product in maintaining viability of PDAC cells.

Gene Symbol	Gene name	Reference	Function	Cancer type
K-ras	v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog	[Fleming et al., 2005]	Ras proteins bind GDP/GTP and possess intrinsic GTPase activity	Pancreatic, colorectal, lung, thyroid, acute myelogenous leukaemia
IGFBP4	Insulin-like growth factor binding protein 4	[Culouscou and Shoyab, 1991]	Insulin-like growth factor binding	Colon cancer
STAT1	Signal transducer and activator of transcription 1, 91kDa	[Sahin et al., 2003]	Mediates signaling by interferons(IFNs)	Leukaemia
SMAD4	SMAD, mothers against DPP homolog 4	[Vogelstein and Kinzler, 2004, Futreal et al., 2004]	Common mediator of signal transduction by TGF-beta (transforming growth factor) superfamily, acts as a tumour suppressor	Colorectal cancer
TFPI2	Tissue factor pathway inhibitor 2	[Sato et al., 2005, Sova et al., 2006]	May play a role in the regulation of plasmin-mediated matrix remodeling. Inhibits trypsin, plasmin, factor VIIa/tissue factor and weakly factor Xa	Non-small-cell lung cancer, pancreatic cancer
TMPRSS4	Transmembrane protease, serine 4	[Kebebew et al., 2005]	Probable protease. Seems to be capable of activating ENaC	Thyroid cancer
TMPRSS2	Transmembrane protease, serine 2	[Mertz et al., 2007]	Serine-type endopeptidase activity	Prostate cancer
SERPINI2	serpin peptidase inhibitor, clade I (pancpin), member 2	[Ozaki et al., 1998]	Protease inhibitor	Breast cancer,pancreatic cancer
CD44	CD44 antigen	[Coppola, 2000]	Involved in cell proliferation, differentiation, migration, and angiogenesis, presentation of cytokines, chemokines, and growth factors to the corresponding receptors	Breast cancer, prostate cancer
FOXM1	Forkhead box M1	[Katoh and Katoh, 2004]	Transcriptional activatory factor	Pancreatic cancer and basal cell carcinoma
BRCA1	Familial breast/ovarian cancer gene 1	[Hezel et al., 2006b]	Facilitating cellular response to DNA repair	Breast/ovarian cancer

Gene Symbol	Gene name	Reference	Function	Cancer type
STK11	Seine/threonine kinase 11	[Sahin et al., 2003]	A tumour suppressor gene	Non small cell lung cancer, lost in Pancreatic cancer
KIF20A	Kinesin family member 20A	[Taniuchi et al., 2005]	A motor required for the retrograde RAB6 regulated transport of Golgi membranes and associated vesicles along microtubules	Pancreatic cancer
ALP	Alkaline phosphate	[Pospisil et al., 2006]	Responsible for removing phosphate groups from many types of molecules, including nucleotides, proteins, and alkaloids	Various cancers
PAP	Prostatic acid phosphatase		Functions as a neutral protein tyrosine phosphatase (PTP) in prostate cancer cells	Prostatic cancer
INHBA	Inhibin, beta A	[Cao et al., 2004]	Inhibins inhibit the secretion of follitropin by the pituitary gland	Ovarian cancer
AKAP12	A-KINASE ANCHOR PROTEIN 12		Anchoring protein that mediates the subcellular compartmentation of protein kinase a (pka) and protein kinase c (pkc)	Pancreatic, prostate, breast and gastric cancer
ELK3	ELK3, ETS-domain protein		Negative regulation of transcription and activate transcription when co-expressed with ras, src or mos	Pancreatic cancer
FOXQ1	Forkhead box Q1		Transcription factor activity	Pancreatic cancer
EIF5A2	eukaryotic translation initiation factor 5A2		Pancreatic, ovarian and colorectal cancer	
EFNA5	Ephrin-A5		Induces compartmentalised signaling within a caveolae-like membrane microdomain when bound to the extracellular domain of its cognate receptor	Pancreatic cancer
KLK6	Kallikrein 6	Hustinx et al. [2004]	Serine-type endopeptidase activity	Pancreatic cancer
CD74	CD74 antigen		MHC class II protein binding	Pancreatic cancer
KLK10	Kallikrein 6		Has a tumour-suppressor role for NES1 in breast and prostate cancer	Pancreatic, breast and prostate cancer
TMPRSS3	Transmembrane protease, serine 4		Serine-type endopeptidase activity	Pancreatic cancer
Muc4	Mucin 4, cell surface associated		ErbB-2 class receptor binding	Pancreatic cancer

Gene Symbol	Gene name	Reference	Function	Cancer type
DMBT1	Deleted in malignant brain tumours 1		May be considered as a candidate tumour suppressor gene for a number of cancers	Pancreatic, brain, lung, esophageal, gastric, and colorectal cancers
CXCL2	Chemokine (C-X-C motif) ligand 2		Produced by activated monocytes and neutrophils and expressed at sites of inflammation	Pancreatic cancer
CEB1	Cyclin-E-binding protein 1		acid-amino acid ligase activity	Pancreatic cancer

Table 2.4: A list of a number of known cancer genes. The list has been annotated with information concerning references, gene function and cancer types in which mutations are found.

2.3.3 Modelling and reasoning over molecular networks

Analysing complex interaction networks is harder than producing them. These networks tend to be very complicated, with thousands of nodes and edges to be considered. For example, the Boehringer Mannheim chart that consists of a complex, interconnecting metabolic processes network, has over 2200 reactions and 820 enzymes see Figure 2.6. Biomedical reaction networks are the subject of extensive modelling studies, they are often modeled by means Differential Equations. Nathalie Chabrier-Rivier proposed a kinetic model of biological pathways using Ordinary Differential Equations (ODEs). Similarly, in [Hurlebaus et al., 2002] the authors used ODEs in modelling the kinetics of the pathways where all real metabolites concentration are considered. Some other recent studies also used ODEs to evaluate their models [Li et al., 2008, Chen et al., 2004]. Despite their expressive power, they are difficult to reason about and make decisions because of the effects that the uncertainty on data may cause. In [Cruz and Barahona, 2005], the authors proposed a constraint reasoning framework to enable safe decision support despite data uncertainty and illustrate the approach in the tuning of drug design. Rule based models are also used for the same purpose. Pathwaylogic, which uses algebraic syntax, is a work presented by Eker et al. [2002]. In [Kim and Park, 2005] the authors present a formalism to represent and analyse protein-protein interaction networks, and a computation tree Logic (CTL) is used as a language to query their system. In [Hvidsten et al., 2003] the authors present a supervised learning approach to predict biological process from gene expression data and biological knowledge, using GO, and then generate hypothesis for unknown genes. BIOCHAM the biomedical abstract machine [Fages et al., 2005] offers automated reasoning tools for querying the temporal properties of a network system, based on a formal semantics to bimolecular interaction maps. The machine learning system of BIOCHAM allows researchers to discover interaction rules from a partial model with constraints on the system behaviour expressed in temporal logic [Calzone et al., 2005]. Some other approachs include BioNetGen [Blinov et al., 2004], Bio-ambients [Regev et al., 2004], Hybrid Petri Nets [Hofestadt and Thelen, 1998], and Hybrid Concurrent Constraint languages [Bockmayr and Courtois, 2002]. In [Tamaddoni-Nezhad et al., 2004], the authors used a logic-based representation and a com-

Modelling method	Reference	Features	consider kinetics	rule-based	High level representation
Ordinary Differential Equations (ODEs)	Li et al. [2008], Chen et al. [2004]	Model kinetic of biological pathways.	✓	×	×
Pathwaylogic	Eker et al. [2002]	Rule-based approach that uses algebraic syntax.	✓	✓	✓
GOEx	Hvidsten et al. [2003]	Supervised learning approach to predict biological process using GO	✓	×	×
BIOCHAM	Fages et al. [2005]	Automated reasoning tool for querying the temporal properties of bimolecular interaction maps.	✓	✓	×
BioNet-Gen	Blinov et al. [2004]	Rule-based modeling of signal transduction based on the interactions of molecular domains.	✓	✓	×
Bio-ambients	Regev et al. [2004]	A calculus used for simulations that provides frame works for molecular and cellular compartmentalisation.	×	✓	✓
Hybrid Petri Nets	Hofestadt and Thelen [1998]	Simulates quantitative modeling of biochemical networks.	×	×	✓
Hybrid Concurrent Constraint languages	Bockmayr and Courtois [2002]	Declarative compositional programming language with a well-defined semantics to model and simulate the dynamics of hybrid systems, which exhibit both discrete and continuous change.	✓	✓	×
Abduction and Induction	Tamaddoni-Nezhad et al. [2004]	Logic-based representation and a combination of Abduction and Induction to model inhibition in metabolic networks.	×	✓	✓

Table 2.5: Modelling biological networks methods

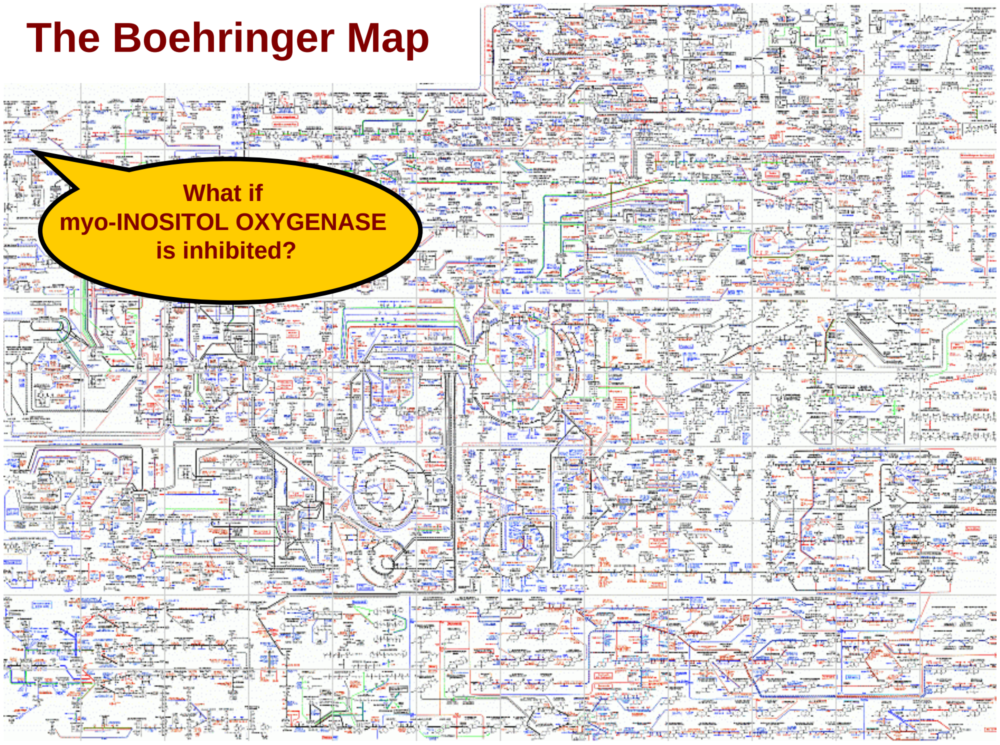


Figure 2.6: The Boehringer Mannheim chart is an example of a condensed metabolic network where manual investigation becomes almost impossible. Automated reasoning over such networks can answer question such as ”What if an enzyme got inhibited in one part of the map ?” What effects does this have on the rest of the map and how can we explain this inhibition?

bination of Abduction and Induction to model inhibition in metabolic networks. The work in [Tamaddoni-Nezhad et al., 2004] serves as a base for the approach that is discussed in this thesis. DRUM [Nejdl and Giefer, 1994] is a belief revision system, it used systems descriptions and observations as model consistency checks. It extend the IMMORTAL [Chou and Winslett, 1991] system by more sophisticated instantiation and inconsistency checking strategies. DRUM uses a combination of static and dynamic inconsistency detection methods and it is able to solve larger and more complicated examples than IMMORTAL. Both DRUM and IMMORTAL were not used in biological networks modelling before.

Chapter 3

From gene expression to interaction networks

In this chapter, the work flow of constructing predicted protein-protein interaction networks starting from a gene expression data set is discussed. The use of structural information to construct such interaction networks which in turn facilitate gene expression data analysis is then motivated. A summary of all the databases utilised in this work is also introduced.

Innovations in experimental methods such as microarray, have enabled large scale analysis of gene expression data, hence provided a foundation for cancer research. The huge amount of data produced by such techniques, stands on the way of a significant analysis of the data in order to use it for diagnostic, prognostic or therapeutic purposes. Protein interactions provide an important context for understanding proteins functions. In $\sim 60\%$ of protein-protein interactions the two interacting proteins share functional similarity, therefore identifying protein interactions is an important component of functional annotation. Further investigations of such data is hampered by the fact that except for the sequence rather little is known about those genes. Therefore, to interpret the data, there is a need for finding the relation between its elements. Large-scale protein interaction maps provide a new global perspective with which to analyse protein function. The main idea of this approach is the prediction of protein interactions using structural data is shown in Figure 3.1. Starting from a set of gene expression data we consider all pairs of sequences for the list. Since structure recognition is still lagging way behind the out put of protein sequence data. To overcome the gap between the known genes sequences and structures, we use GTD [McGuffin and Jones, 2003a], a database that applies threading to predict the structure of all proteins with unknown structures. Our interaction prediction approach is based on SCOPPI, the Structural Classification of Protein-Protein interaction [Winter et al., 2006b], that contains $\sim$ 100.000 interactions of all the structurally observed interactions between protein domains. It computes domain-domain interactions for all multi-domain and multi-chain proteins in the Protein Data Bank (PDB) [Berman et al., 2000]. The method then queries SCOPPI for a an interaction template between two domains, then two proteins are considered to be interacting if they contain two domains and SCOPPI contains an interaction between these two domains Another major challenge that faces gene expression data analysis is that most of the information is hidden in a huge amount of publications.

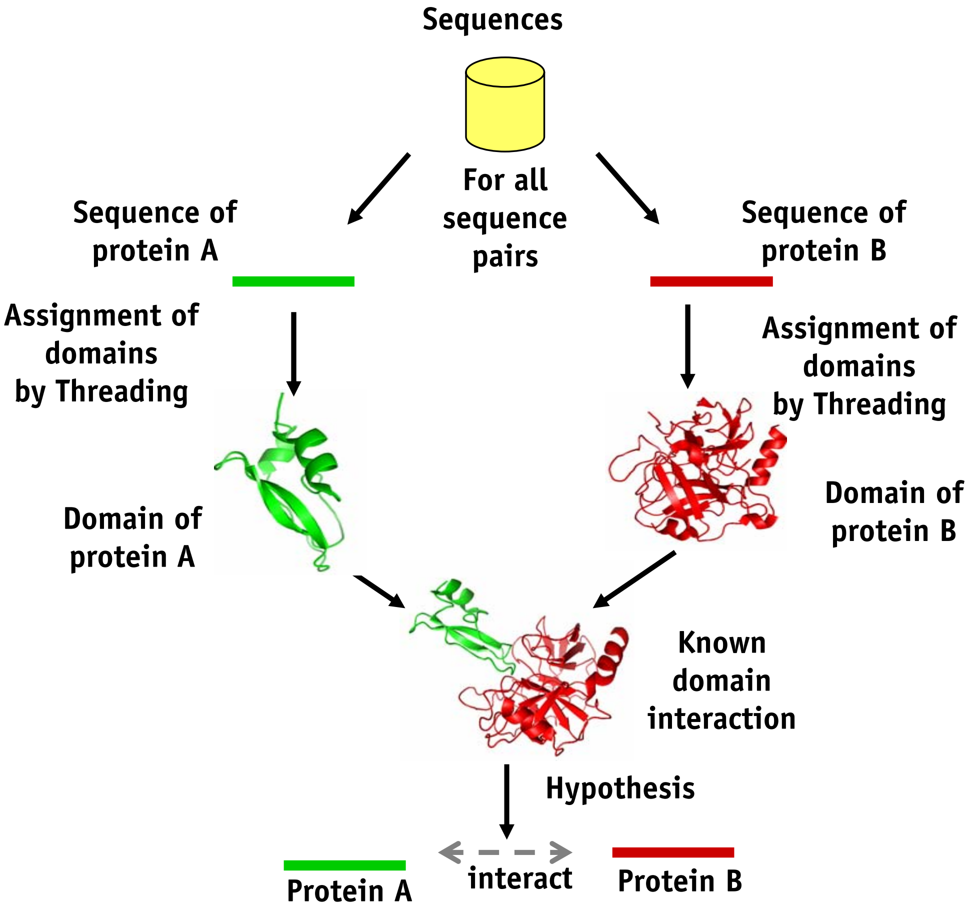


Figure 3.1: From sequences to protein interactions: For all pairs of sequences of unknown structure we use threading for the assignment of domains. Two proteins are considered to be interacting if they contain two domains and SCOPPI contains an interaction between these two domains

3.1 Data sources

In this section the work flow linking gene expression data, biological pathways and interaction data is described. The underlying data sources used are briefly summarised.

3.1.1 3D structures - PDB

The Protein Data Bank (PDB) [Berman et al., 2000] is a repository of information for 3D structures of large biological molecules, including proteins and nucleic acids. As of July 2008, it contains some 51860 protein structures of which 44,000 have been obtained by X-ray crystallography, 180 by electron microscopy and 7300 by NMR. Around half of the PDB structures are multi-domain structures.

3.1.2 Classification of Proteins - SCOP

The structural classification of proteins (SCOP), is a hierarchical classification of protein structures at domain level. The hierarchy contains four levels (class, fold, superfamily, family). At the family level domains share a high sequence similarity and hence are structurally very similar. At superfamily level there is still good structural agreement concerning the overall topology despite possibly low sequence similarity. Domains grouped at family and superfamily level can be considered homologous. Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement with the same topological connections. Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they composed: (1) all alpha (for proteins whose structure is essentially formed by alpha-helices), (2) all beta (for those whose structure is essentially formed by beta-sheets), (3) alpha and beta (for proteins with alpha-helices and beta-strands that are largely inter-spersed), (4) alpha plus beta (for those in which alpha helices and beta strands are largely segregated) and (5) multi-domain (for those with domains of different fold and for which no homologues are known at present) [Murzin et al., 1995].

3.1.3 Domain domain interactions - SCOPPI

The Structural Classification of Protein-Protein Interfaces (SCOPPI) is a database containing all domain-domain interactions in the PDB. SCOPPI applies SCOP domain definitions and a distance criterion to determine inter-domain interfaces. Using a novel method based on multiple sequence and structural alignments of SCOP families, SCOPPI presents a comprehensive geometrical classification of domain interfaces. Various interface characteristics such as number, type and position of interacting amino acids, conservation, interface size, and permanent or transient nature of the interaction are further provided. Two domains are considered as interacting if there are at least 5 residue pairs within 5Å [Winter et al., 2006b].

3.1.4 Threading - GTD

Threading is a computational method for protein structure prediction from amino acid sequence regardless of any sequence similarity, since dissimilar sequences can still adopt similar folds.

The basic idea is that the target sequence (the protein sequence for which the structure is being predicted) is threaded or fitted through the backbone structures of a library of template of known protein structure to find out which one is most compatible as measured by a "goodness of fit" score calculated for each sequence-structure alignment. Threading is more sensitive than sequence alignment and can still assign folds correctly despite low similarity. I utilise the Genomic Threading Database (GTD) [McGuffin and Jones, 2003b] which contains structural folds assignments to proteins with unknown structure. Annotations are based on GenTHREADER [McGuffin and Jones, 2003a], a reliable fold recognition method. In the GTD 84% of the sequences are assigned at a p-value < 1 (certain- medium) confidence.

3.1.5 Gene ontology annotations - GO

The Gene Ontology [Ashburner et al., 2000] is a controlled vocabulary to describe gene and gene product attributes in any organism. These attribute covers three categories: molecular function, biological process and cellular component of gene products. The biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation. Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition also applies to the capability that a gene product (or gene product complex) carries as a potential. It describes only what is done without specifying where or when the event actually occurs. Cellular component refers to the location in the cell where a gene product is active.

3.1.6 Confirmed interactions - NetPro, HPRD, BIND

NetPro. NetPro[1] is the proprietary protein interaction database covering more than 200,000 expert curated and annotated of Protein-Protein, Protein-Small molecules DNA and RNA interactions. All the interactions are extracted from peer reviewed published scientific literature by semiautomated method and then have gone through significant quality checks in terms of expert cross-checking.

The Human Protein Reference Database - HPRD HPRD[2] contains annotations for more than 2750 human proteins. Apart from interaction annotations, it additionally includes annotations for post-translational modifications, enzyme-substrate relationships and disease associations. The database was derived through manual curation by expert biologists interpreting more than 300.000 published articles. Interactions are classified in *in vivo*, *in vitro*, and two-hybrid.

The Biomolecular interaction database - BIND BIND [Bader and Hogue, 2000] includes high-throughput experimental data as well as complexes from PDB. BIND include different types of interactions such as interactions between any two molecules composed of proteins, nucleic acids and small molecules. It also describes chemical reactions, photochemical activation and conformational changes.

[1] https://www.molecularconnections.com/protein_interactions.html

[2] www.hprd.org/

3.1.7 Biological Pathways - KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a resource that provides a reference knowledge base for linking genomes to biological systems and wiring diagrams of interaction networks and reaction networks known as the KEGG Pathways database [Kanehisa et al., 2002]. The KEGG Pathway database is a collection of manually drawn pathway maps for metabolism, genetic information processing, environmental information processing such as signal transduction, various other cellular processes and human diseases.

3.2 Interaction network construction

3.2.1 Algorithm and implementation

A summary of the work flow is presented by the pseudo-code in Table 3.1

3.2.1.1 Structure-based prediction of protein interactions.

We implemented a methodology that utilises structural data from SCOPPI to predict potential interaction within a gene expression data set. The resulting potential interactions are further investigated by considering amino acid sequence conservation of $\geq 50\%$ at the interaction interface when compared to the structural template. In the following we describe the working steps of the method as shown in Figure 3.2

3.2.1.2 Gene expression data of disease relevant genes

Initially, a list of genes obtained from a disease microarray data set is used. The genes are usually clustered into lists that identify the genes as "up" or "down" regulated. The genes are supplied as lists of Affymetrix[3] ids which we then map to their corresponding proteins Ensemble[4] ids.

3.2.1.3 From genes to structure assignment and family classification.

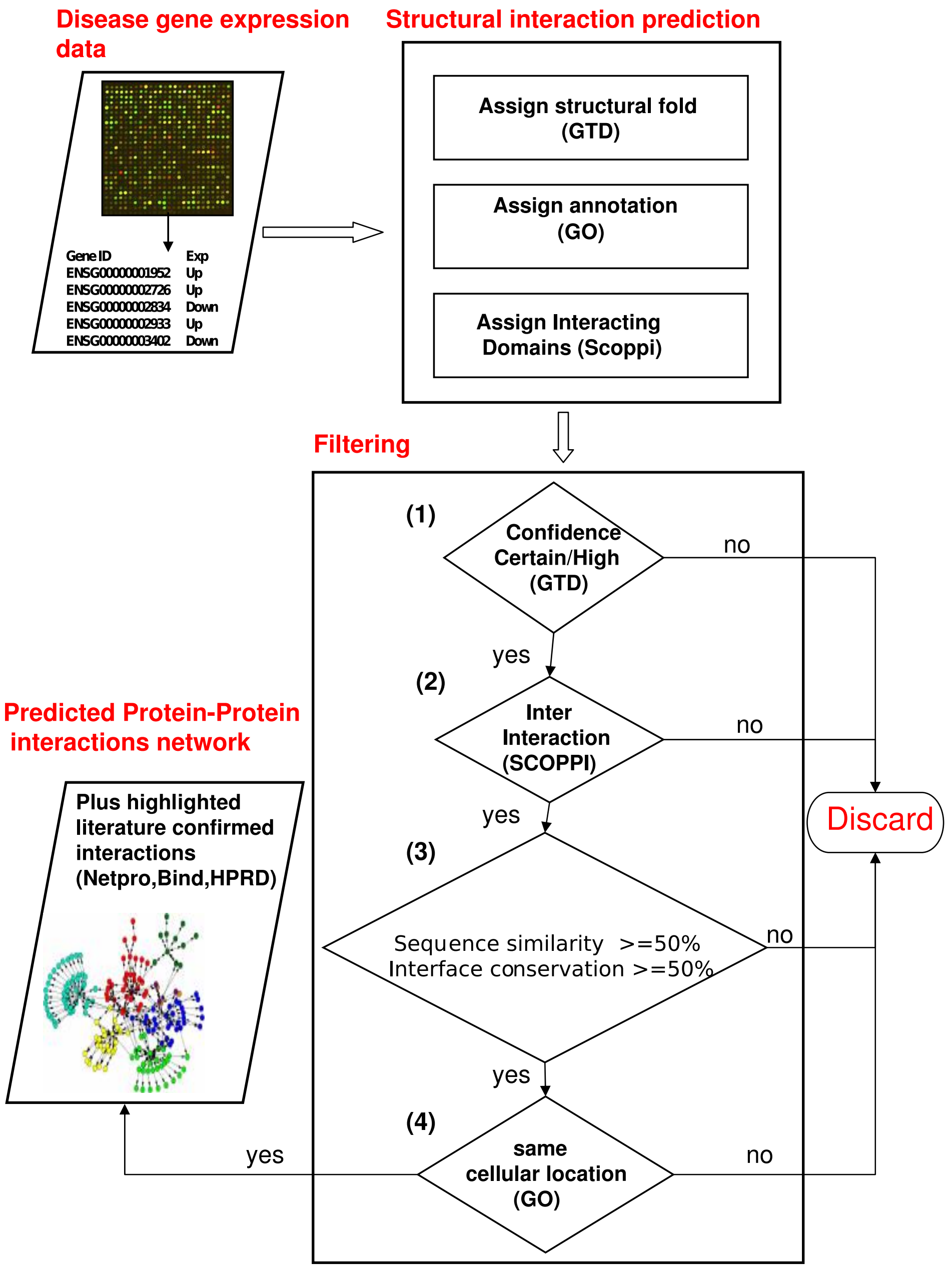
Only when proteins fold into one, or more, specific spatial conformations are they able to perform their biological function. Most of the genes from gene expression data sets are of unknown structure. First, the Genomic Threading Database (GTD) as fold recognition method to assign SCOP structural families to the proteins in our data sets is used. Only assignments with certain and high confidence by GTD are considered.

3.2.1.4 From structural folds to domain interactions.

For the assigned SCOP domains, SCOPPI is used to identify interacting domain pairs. In this step, two proteins are considered as interacting if each contains a domain where there is structural evidence for such a domain–domain interaction according to SCOPPI. The evidence in-

[3] http://www.affymetrix.com/index.affx

[4] http://www.ensembl.org/index.html


Disease gene expression data
Structural interaction prediction
Assign structural fold (GTD)
Assign annotation (GO)
Assign Interacting Domains (Scoppi)
Filtering
(1)
Confidence Certain/High (GTD)
no
yes
(2)
Inter Interaction (SCOPPI)
no
yes
(3)
Sequence similarity >=50%
Interface conservation >=50%
no
yes
(4)
same cellular location (GO)
no
yes
Discard
Predicted Protein-Protein interactions network
Plus highlighted literature confirmed interactions (Netpro,Bind,HPRD)

Gene ID	Exp
ENSG00000001952	Up
ENSG00000002726	Up
ENSG00000002834	Down
ENSG00000002933	Up
ENSG00000003402	Down

Figure 3.2: The work flow illustrating the interactions prediction method different steps: Starting with experimental data such as microarray, is provided as a list of genes grouped into up/down regulated clusters. Structural interaction prediction: to obtain the threaded PDB structures and SCOP family assignments of the genes we use the GTD database, the genes are annotated using the Gene Ontology (GO), the SCOPPI database is then queried for obtaining the domain interactions among the genes. Filtering: the filtering criterion used are: Only folds with certain and high confidence assignments, domain interactions between domains from different polypeptide chains (intra interactions), interface conservation $\geq$50% and interactions between proteins located at the same cellular location are considered to screen out the highly confident interactions. Output:a protein-protein interaction network where known interactions are then identified by checking against the experimentally confirmed interaction databases (NetPro, HPRD and BIND)

```
differentiallyExpressedGenes = MicroArray(PancreasTissue)
populate PPI-network with known interactions
for gene in differentiallyExpressedGenes:
  assign structural fold from GTD to gene -> gene.structuralFold
  assign GO annotation to gene -> gene.GO

for gene1 in differentiallyExpressedGenes:
  for gene2 in differentiallyExpressedGenes:
    if gene1.structuralFold and gene2.structuralFold    and
      have structural Template in SCOPPI:
      foldConfidence = confidence(gene1.structuralFold) and
      confidence(gene2.structuralFold)                        (1)

      templates = SCOPPI(gene1.structuralFold, gene2.structuralFold)

      interInteraction = False
      acceptableInterfaceModel = False

      for template in templates:                               (2)
        if template is interaction-Type inter:
           interInteraction = True

        alignment1 = align(template.sequence1, gene1.sequence) (3)
        alignment2 = align(template.sequence2, gene2.sequence)

        if sequenceSimilarity(alignment1) >= 50%   and
         sequenceSimilarity(alignment2) >= 50%     and
         interfaceConservation(alignment1) >= 50%  and
         interfaceConservation(alignment2) >= 50%  :
          acceptableInterfaceModel = True

      sameLocation = (intersection(gene1.GO.location,           (4)
                     gene2.GO.location) is non-empty)

      if foldConfidence         and
       interInteraction         and
       acceptableInterfaceModel and
       sameLocation:
        add (gene1, gene2) to PPI-network
```

Table 3.1: Algorithm pseudo-code. The numbering to the left refers to numbering in of the different steps in Figure 3.2

teraction then serves as a structural template to model the predicted interaction. Figure 3.1 sketches the structure assignment and interaction prediction steps of the method. This initial predicted interaction network is then further refined.

3.2.1.5 Refinement of predicted Protein-protein interactions

Interface conservation evaluation It has been shown that protein interface residues are usually more conserved than the rest of the exposed surface [Elcock and McCammon, 2001, Valdar and Thornton, 2001]). In order to compute the interface conservation, the information about residues in the interface is taken from the SCOPPI database, an interface consists of all atoms and residues of a domain that are within 5 $\AA$ of another domain. We align the original protein sequence against the SCOPPI template sequence and calculate the sequence identity percentage of the interface residues. The evaluation criterion is explained as follows: If one protein has a conservation of more than 50 % of residues at interface against counterpart of the known template structure, we assume that they share the same interaction partner. For interesting examples, we perform structural alignment on their predicted interacting pairs with their corresponding SCOPPI templates. We finally consider pairs that are aligned with an RMSD $< 2\AA$. For examples of interest, we perform further investigation of the interface, where we check for the presence of key residues and conservation of the active and binding sites. Information about key residues is mainly extrated manually from literature and to query for active and binding sites we use swissprot - a manually curated protein sequence database which provides a high level of annotation that includes functions of the proteins, post-translational modifications, domains and binding and active sites, secondary structure, diseases associated with deficiencies in proteins[5].

False positive reduction in the protein-protein interaction prediction using GO cellular location All proteins are annotated with their corresponding GO terms. To reduce false positives from the predicted interaction network, we screen out interactions at this stage with one partner annotated as exclusively intra- and the other as exclusively extra-cellular.

Inter and Intra interactions The scoppi interactions are classified into four groups: (i) homo-intra, (ii) homo-inter, (iii) hetero-intra, and (iv) hetero-inter. Homo- or hetero- is assigned depending on whether the interacting domains are from the same family or from different families, respectively [Kim et al., 2006]. *Intra*-interactions are assigned to domain pairs from the same chain, where one chain forms two domains which then interact. *Inter*-interactions take place between domain pairs from different chains depeicted mainly in complex structures. Since we are mianly interested in interactions between diffrent proteins, we only consider hetero-inter interactions.

3.2.1.6 Literature confirmed intercations

In order to validate the method, the refined network of predicted protein-protein interactions is compared to those confirmed by experimental interaction databases. For this purpose, NetPro, BIND [Bader and Hogue, 2000], and HPRD [Peri et al., 2003] are used.

[5] http://www.expasy.ch/sprot/

3.2.1.7 From gene expression data to Pathways

In the pathway analysis approach. The aim is to construct a PDAC related pathway network that resembles the regulatory circuits which are disrupted in the cell. To this end, the KEGG Pathways database is queried, genes are then grouped according to the pathways they are involved in. We define two pathways to be related if they share at least four genes. Finally, we obtain an overview of the related pathways which are mainly modulated in PDAC. It can help in understanding the processes the pancreas cell undertakes to become malignant.

Chapter 4

Four case studies using structural templates to predict novel protein interactions and targets from pancreas tumour gene expression data

4.1 Case study I: A novel pancreatic cancer network of known and predicted protein-protein interactions

4.1.1 Introduction

A number of gene expression screens have been carried out to identify genes differentially expressed in cancerous tissue. To identify molecular markers and suitable targets, these genes have been mapped to protein interactions to gain an understanding at systems level. In this chapter, a detailed description of four pancreatic cancer case studies are presented. We analysed the data using protein interactions prediction methods to the four data sets of pancreatic cancer gene expression data. I will guide the reader through the single steps towards constructing the protein interaction network. The evaluation steps are discussed and then examples are discussed in details demonstrating the success of our method in highlighting genes to be considered for laboratory experimental as potential drugs or diagnostic markers.

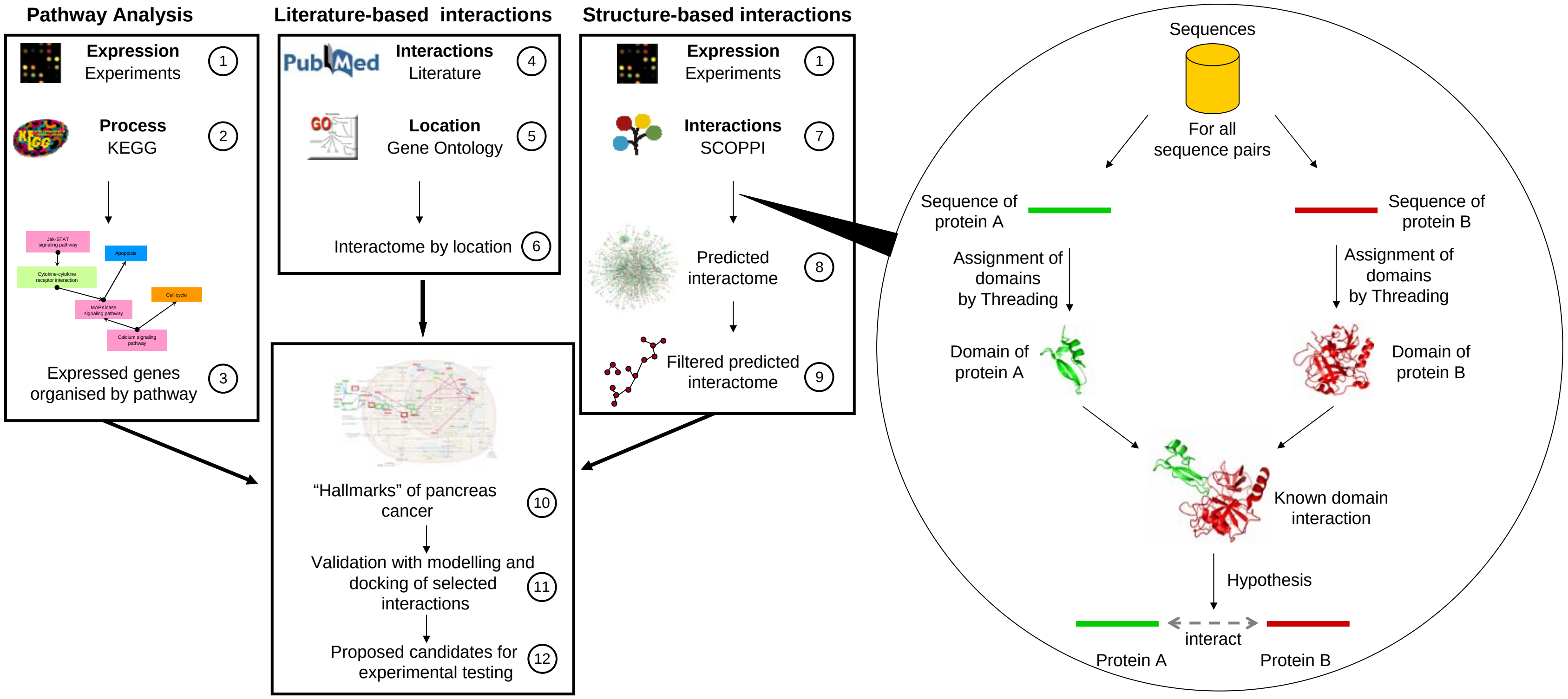


Figure 4.1: Approach of this study. We start with the PDAC gene expression data (1). Using KEGG a pathways database(2), Interactions from literature, SCOPPI(7), Gene Ontology annotation (5), and interaction predictions (8,9), we construct views highlighting different aspects of the gene data set (3,6,9). These are then integrated into a comprehensive interaction map (10) which is then overlayed with the pathways of cancer hallmarks [Hanahan and Weinberg, 2000]. Finally, promising candidates as therapeutic markers are identified and validated by computational methods (11) in order to provide targets for experimental testing (12).

4.1.2 Approach

In this study, we applied the computational approach described in chapter 3 to automatically reconstruct interaction networks from genes involved in pancreatic cancer, we also obtain a map of pathway alterations and key interactions. We compare this map to the "Hallmarks of cancer" diagram published by [Hanahan and Weinberg, 2000]. The overview of the approach is illustrated in Figure 4.1.

4.1.2.1 Data set: Gene expression data

Our collaborators from the University Hospital, Technical University of Dresden, obtained nineteen tissue samples from surgical specimens from patients who were treated at the Department of Visceral-, Thoracic- and Vascular Surgery, University Hospital Carl Gustav Carus, Technical University of Dresden and the Department of General Surgery and Thoracic Surgery, University of Kiel between 1996 and 2003. Normal pancreatic tissue was obtained from 13 patients who underwent pancreatic resection for other pancreatic diseases. PDAC cells and normal ductal cells were microdissected manually. Microdissection is used in their study, because it has the advantage that the isolated RNA of tumour cells is not less contaminated by RNA from other cell-types.

Our data set (Figure 4.1 (1)) used in this case study originates from four microarray studies performed in the above samples. The data was obtained by integrating various analyses of the gene expression profiles of PDAC from Affymetrix GeneChip experiments such as microdissection, systematic isolation of genes [Grützmann et al., 2003, Grutzmann et al., 2003, 2004a], and the meta-analysis of PDAC gene expression profiles from publicly available data [Grutzmann et al., 2005]. These studies compare expression profiles of pancreatic ductal adenocarcinoma cells to healthy exocrine pancreas cells and only genes which have a fold change of > 2 compared to healthy pancreas tissue are considered. The data set pooled from these studies contains 1612 genes differentially expressed in pancreatic ductal adenocarcinoma (PDAC). Around 1,500 of the genes in our data set are validated by checking them against previously reported differentially expressed cancer genes from [Higgins et al., 2007].

4.1.2.2 From expression to pathways

Our first approach is the construction of a PDAC related pathway network that resembles the regulatory circuits which are disrupted in the cell (3). To this end, we check in which KEGG pathways (2) our dataset genes participate. We query the KEGG Pathways database, genes are then grouped according to the pathways they are involved in. We define two pathways to be related if they share at least four genes. The resulting model is shown in Figure 4.2. We obtain an overview of the related pathways which are mainly modulated in PDAC. It can help in understanding the processes the pancreas cell undertakes to become malignant.

4.1.2.3 Known interactome by localisation

We obtain all experimentally known interactions within our data set from the literature (4) by help of the NetPro database. Within the proteins of the data set 1121 interactions were found in

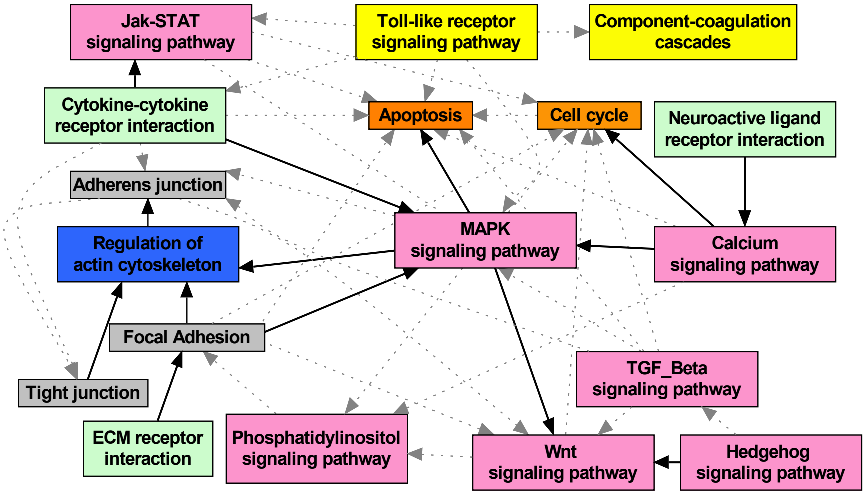

Jak-STAT signaling pathway
Toll-like receptor signaling pathway
Component-coagulation cascades
Cytokine-cytokine receptor interaction
Apoptosis
Cell cycle
Neuroactive ligand receptor interaction
Adherens junction
Regulation of actin cytoskeleton
MAPK signaling pathway
Calcium signaling pathway
Focal Adhesion
Tight junction
TGF_Beta signaling pathway
ECM receptor interaction
Phosphatidylinositol signaling pathway
Wnt signaling pathway
Hedgehog signaling pathway

Figure 4.2: Overview of related pathways that are mainly affected and modulated in pancreatic ductal adenocarcinoma (PDAC). Pathways are grouped according to their similar functions, and each group is coloured differently (pink for signal transduction, yellow for immune system, orange for cell growth and death, light green for signalling molecules and interaction, blue for cell motility, and grey for cell communication). Solid arrows indicate that two pathway have at least four genes in common. Dashed arrows indicate that one pathway is downstream of another according to KEGG (see text for details).

NetPro. We then retrieve localisation information from the Gene Ontology cellular component annotation (5) for every protein. From this, we construct an integrative map of known pancreatic cancer relevant protein interactions (6).

4.1.2.4 Structure-based interaction predictions

Protein interactions provide an important context for understanding protein function. We use structural information to predict novel interactions among the PDAC proteins which can functionally annotate uncharacterised cancer related genes. Our interaction prediction approach is based on SCOPPI, the Structural Classification of Protein–Protein Interfaces [Winter et al., 2006b]. The idea of predicting new interactions from these known ones is sketched in Figure 4.1 on the right (see Chapter 2 for details). The resulting set of initial interaction predictions (Figure 4.1 (8)) yields $\sim$ 700 potential interactions among the PDAC microarray data set. Filtering out predictions with less than 50% interface identity and medium or low GTD confidence results in a set of 84 confident, novel interactions. Table 4.1 contains all the predicted interactions in addition to two literature confirmed interactions.

4.1.2.5 A pancreatic cancer map

By linking the pathway approach, known interactions and structure-based interaction predictions, we produce a detailed PDAC cell map (10). The map illustrates the gene products of the

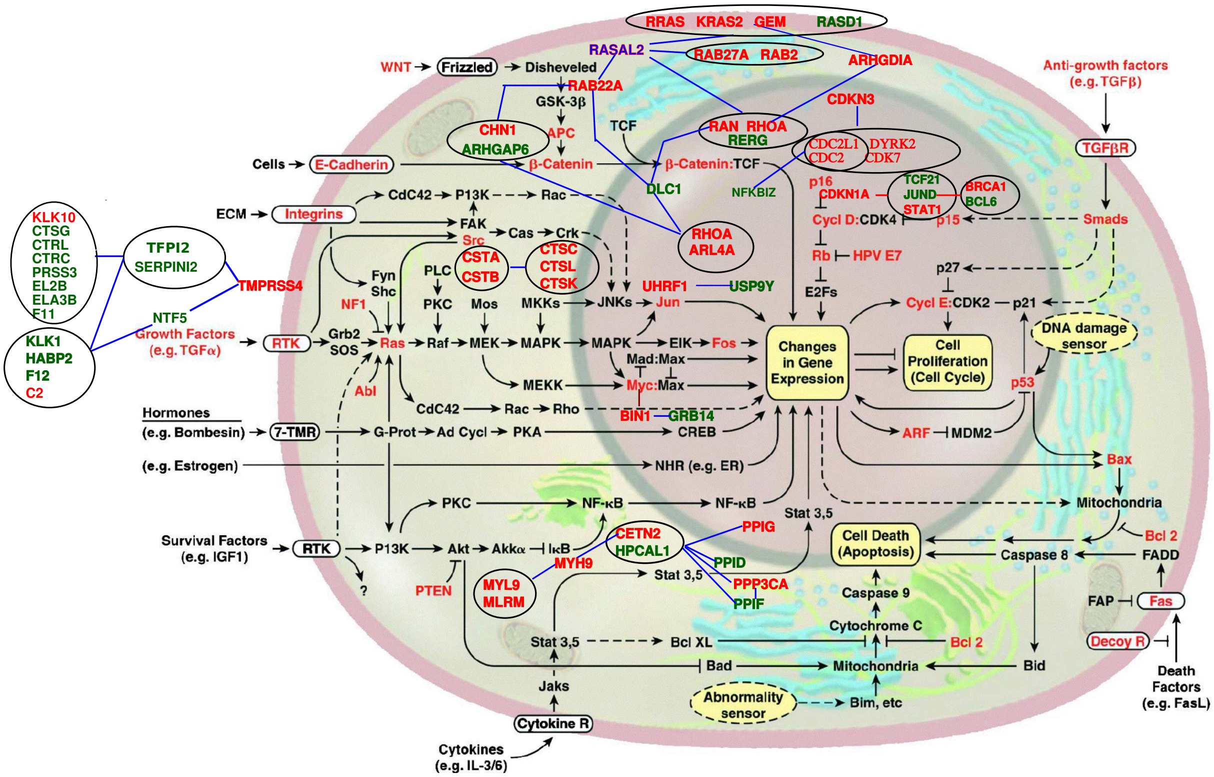
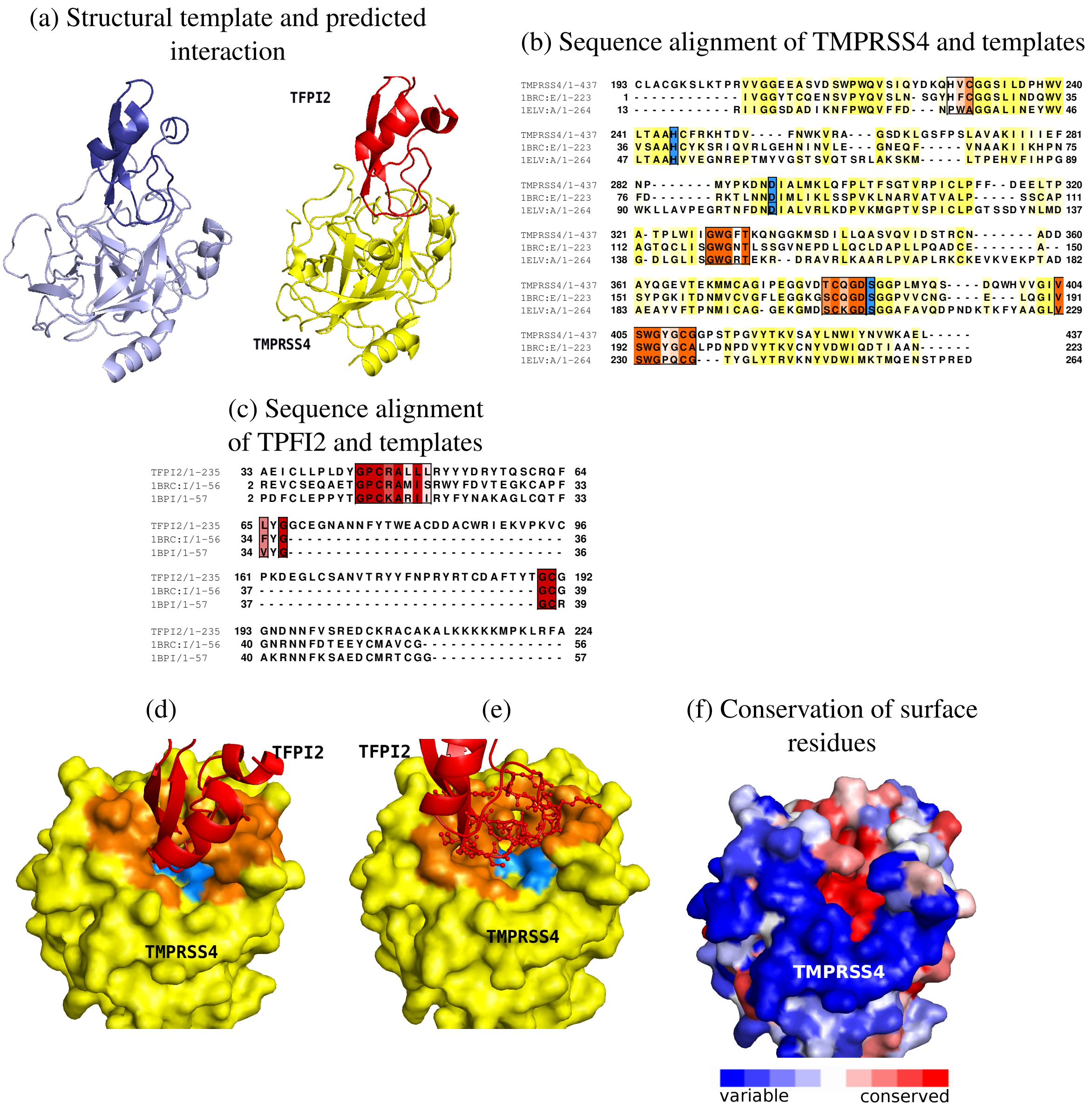


Figure 4.3: A comprehensive map of pancreatic cancer relevant interactions and pathways. The underlying picture was taken from [Hanahan and Weinberg, 2000] and updated with our findings. It depicts an integrated circuit of the cell progress annotated with the PDAC genes that are involved in the novel predicted interactions. Proteins are shown according to their cellular localisation and their associated pathways. Genes coloured green are downregulated while genes coloured red are upregulated. Lines linking genes represent interactions among the genes. Interactions confirmed by literature are indicated by red lines, and predicted interactions are indicated by blue lines.

PDAC data that are involved in all novel predicted interactions, see Figure 4.3. For a better visualisation of the interaction map, we use an edge reduction representation, where the edge connecting two circles indicates that all the elements of one circle interact with all the elements of the other.

4.1.3 Results and Discussion

An interesting example is the role of the downregulated tissue factor pathway inhibitor 2 (TFPI2) as a potential inhibitor of the upregulated transmembrane protease, serine 4 (TMPRSS4). This example is elaborated and discussed in the following section.



(a) Structural template and predicted interaction

(b) Sequence alignment of TMPRSS4 and templates

(c) Sequence alignment of TPFI2 and templates

(d)

(e)

(f) Conservation of surface residues

TFPI2
TMPRSS4
TMPRSS4/1-437 193 CLACGKSLKTPRVVGGEEASVDSWPWQVSIQYDKQHVCGGSILDPHWV 240
1BRC:E/1-223 1 -------------IVGGYTCQENSVPYQVSLN-SGYHFCGGSLINDQWV 35
1ELV:A/1-264 13 ------------RIIGGSDADIKNFPWQVFFD---NPWAGGALINEYWV 46
TMPRSS4/1-437 241 LTAAHCFRKHTDV----FNWKVRA---GSDKLGSFPSLAVAKIIIIEF 281
1BRC:E/1-223 36 VSAAHCYKSRIQVRLGEHNINVLE---GNEQF-----VNAAKIIKHPN 75
1ELV:A/1-264 47 LTAAHVVEGNREPTMYVGSTSVQTSRLAKSKM-----LTPEHVFIHPG 89
TMPRSS4/1-437 282 NP-------MYPKDNDIALMKLQFPLTFSGTVRPICLPFF--DEELTP 320
1BRC:E/1-223 76 FD-------RKTLNNDIMLIKLSSPVKLNARVATVALP-----SSCAP 111
1ELV:A/1-264 90 WKLLAVPEGRTNFDNDIALVRLKDPVKMGPTVSPICLPGTSSDYNLMD 137
TMPRSS4/1-437 321 A-TPLWIIGWGFTKQNGGKMSDILLQASVQVIDSTRCN-------ADD 360
1BRC:E/1-223 112 AGTQCLISGWGNTLSSGVNEPDLLQCLDAPLLPQADCE-------A-- 150
1ELV:A/1-264 138 G-DLGLISGWGRTEKR--DRAVRLKAARLPVAPLRKCKEVKVEKPTAD 182
TMPRSS4/1-437 361 AYQGEVTEKMMCAGIPEGGVDTCQGDSGGPLMYQS----DQWHVVGIV 404
1BRC:E/1-223 151 SYPGKITDNMVCVGFLEGGKGSCQGDSGGPVVCNG----E---LQGIV 191
1ELV:A/1-264 183 AEAYVFTPNMICAG-GEKGMDSCKGDSGGAFAVQDPNDKTKFYAAGLV 229
TMPRSS4/1-437 405 SWGYGCGGPSTPGVYTKVSAYLNWIYNVWKAEL----- 437
1BRC:E/1-223 192 SWGYGCALPDNPDVYTKVCNYVDWIQDTIAAN------ 223
1ELV:A/1-264 230 SWGPQCG---TYGLYTRVKNYVDWIMKTMQENSTPRED 264
TFPI2/1-235 33 AEICLLPLDYGPCRALLLRYYYDRYTQSCRQF 64
1BRC:I/1-56 2 REVCSEQAETGPCRAMISRWYFDVTEGKCAPF 33
1BPI/1-57 2 PDFCLEPPYTGPCKARIIRYFYNAKAGLCQTF 33
TFPI2/1-235 65 LYGGCEGNANNFYTWEACDDACWRIEKVPKVC 96
1BRC:I/1-56 34 FYG----------------------------- 36
1BPI/1-57 34 VYG----------------------------- 36
TFPI2/1-235 161 PKDEGLCSANVTRYYFNPRYRTCDAFTYTGCG 192
1BRC:I/1-56 37 -----------------------------GCG 39
1BPI/1-57 37 -----------------------------GCR 39
TFPI2/1-235 193 GNDNNFVSREDCKRACAKALKKKKKMPKLRFA 224
1BRC:I/1-56 40 GNRNNFDTEEYCMAVCG--------------- 56
1BPI/1-57 40 AKRNNFKSAEDCMRTCGG-------------- 57
TFPI2
TMPRSS4
TFPI2
TMPRSS4
TMPRSS4
variable
conserved

Figure 4.4: Example for a predicted interaction between transmembrane protease, serine 4 (TMPRSS4) and tissue factor pathway inhibitor 2 (TFPI2) (a) The known complex of trypsin (light blue) and amyloid beta-protein precursor inhibitor (dark blue) serves as a template to predict and model the interaction between TMPRSS4 (yellow) and TFPI2 (red). (b) Alignment of the sequence to model the TMPRSS4 structure and the sequence of the template. Interface residues are shown in orange, the catalytic triad is shown in blue. Sequence similarity is shown in shades of colour. (c) Alignment of the sequence to model the TFPI2 structure and the sequence of the template. Interface residues in red. (d) Close-up view of the predicted interaction of TMPRSS4 and TFPI2. The interface region of TMPRSS is shown in orange, with catalytic triad of the active site shown in blue. (e) After energy minimisation, the pocket slightly opens and initial minor clashes can be resolved. (f) Amino acid conservation colouring of the predicted TMPRSS4 structure shows a well-conserved pocket.

4.1.3.1 Pathways in pancreatic cancer

Comparison of predicted with known cancer pathways.

A number of pathways are known to be affected by PDAC. The Wnt and Hedgehog signalling pathways are essential during embryonic pancreatic development. The misregulation of these pathways has been implicated in several forms of cancer and may also be an important mediator in human pancreatic carcinoma. Thayer et al. and Kayed et al. suggested that these pathways may have an early and critical role in the genesis of this cancer, and that maintenance of the Hedgehog signalling is important for aberrant proliferation and tumorigenesis.

The Notch signalling pathway has been shown to contribute to human cancers when abnormally regulated [Hezel et al., 2006b]. Xu and Attisano presented in [Xu and Attisano, 2000] a study that revealed a mechanism for tumorigenesis whereby genetic defects in SMADs induce their degradation through the ubiquitin-mediated pathway.

The pathways that are affected by the deregulation of genes in pancreatic cancer are shown in Figure 4.2. The analysis of such a network can help to explain how the deregulated pathways affect each other and how this might result in tumorigenesis. Cancerous cells typically affect a variety of cellular pathways that are related to cell growth, cell division, evasion of apoptosis, and signalling [Hanahan and Weinberg, 2000]. Comparing our pathway analysis to these general cancer mechanisms, our results indicate that in pancreatic cancer the calcium signalling pathway is affected. The key function of the exocrine pancreas is to synthesise, package and secrete a variety of digestive enzymes. This process is regulated by neurotransmitters and hormones, both of which utilise calcium as a principal signalling molecule [Yano et al., 2003]. Calcium can mediate signalling transduction by activation of a number of calcium-activated protein kinases and protein phosphatases such as calcineurin [Williams, 2001]. It also plays an important role in primary signalling mechanism control secretion. In addition, we observe that the MAPKinase pathway has the highest connectivity which supports the hypothesis that it plays a crucial role in tumorigenesis. Hedgehog, Wnt and Jak-STAT signalling pathways transduce the signals from the extracellular environment. All together they perturb cell adhesion, cell cycle, and the apoptosis pathway which ultimately leads to the abnormal phenotype of PDAC. Finally, they pave way for invasion and metastasis, enabling cancer cells to escape the primary tumour mass and colonise new terrain in the body.

4.1.3.2 Hallmark interactions of pancreatic cancer

Combining pathways, known interactions and predicted interactions, we obtain the hallmarks of pancreatic cancer map (Figure 4.3). For a better visualisation of the interaction map, we use an edge reduction representation, where the edge connecting two circles indicates that all the elements of one circle interact with all the elements of the other. Our data confirms several of the classical cancer alterations. In addition, we complement these by known and predicted interactions. Most notably, we find many extracellular proteins to be deregulated. Table 4.1 lists all structure-based interaction predictions after filtering. These interactions have a high confidence with respect to the threading structure prediction method. Furthermore, they have a sufficient conservation of the putative interacting residues when compared to the known structural template that was used to model this interaction. One interesting example of two extracellular proteins that might play a major role in tissue infiltration and metastasis of pancreatic cancer is discussed below.

4.1.3.3 TFPI2 is a potential inhibitor of TMPRSS4

The interaction between the upregulated transmembrane protease, serine 4 (TMPRSS4) and the downregulated tissue factor pathway inhibitor 2 (TFPI2) marks an interesting example.

Cancer invasion and metastasis is a complex mechanism that includes a variety of cellular processes, among which the proteolytic degradation of extracellular matrix has been considered one of the most critical events. The matrix degradation can be promoted by the imbalance between proteolytic enzymes (proteases) and their inhibitors In pancreatic cancer cells [Sato et al., 2005]. TMPRSS4 is involved in the process of metastasis formation and tumour invasion, and its expression is correlated with the metastatic potential [Wallrapp et al., 2000]. TFPI2 is an extracellular protein that belongs to the small Kunitz inhibitor family. It is known to be downregulated in PDAC.

Figure 4.4 shows how our structure-based method predicts and models an interaction between TMPRSS4 and TFPI2. The structures are predicted according to the domains found by Threader. Searching the SCOPPI database for interactions of related domains, we find the complex of trypsin (light blue) and amyloid beta-protein precursor inhibitor (dark blue). The modelled structures (red and yellow in Figure 4.4a) are superimposed with the template of known interaction (blue) to model the putative interaction between them. This interaction is shown again from a different angle in Figure 4.4d. TMPRSS4 residues that are part of the interface are coloured orange, and the catalytic triad of serine, aspartate and histidine is coloured blue. After energy minimisation of the complex, the pocket around the active site slightly opens (Figure 4.4e) and minor clashes that were present before disappear. The sequence alignments of TMPRSS4 and TFPI2 with the sequences of their GTD-assigned structures as well as the SCOPPI structural template are shown in Figure 4.4b and c. Sequence similarity is reflected by shades of colour. We find the interface regions (orange/red) to be well conserved.

This interaction could explain the mechanism of metastasis that makes PDAC a very aggressive type of cancer. TFPI2 is an extracellular matrix associated serine protease inhibitor [Rao et al., 1996] that plays a major role in extracellular matrix degradation during tumour cell invasion and metastasis, wound healing, and angiogenesis. It plays a major role in negative regulation of the coagulation cascades (upper right in Figure 4.2) and its downregulation is associated with malignant pancreas tumours. On the other hand, TMPRSS4 is known to be upregulated in pancreatic cancer, which may be of importance for processes involved in metastasis formation and tumour invasion [Wallrapp et al., 2000].

We can thus hypothesise that TFPI2 acts as a natural inhibitor of TMPRSS4. Since TFPI2 is downregulated, the upregulated TMPRSS4 is no longer inhibited and might facilitate tissue invasion. As a result of applying the structural prediction technique, we predicted 81 novel interactions. Most of the interactions are of extracellular genes which again bring to light the important role of the extracellular interaction network that promote metastasis and influence to fast spreading of the PDAC. We compiled the following table that contains the potential interactions among the genes of our data set. We used the same cutoffs criterion as before (50%) sequence similarity between original and predicted structure.

	Protein 1	Description	up/ down	Interface con-served	Complex PDB ID	Protein 2	Description	up/ down	Interface con-served	Confirmed by litera-ture
1	CDC2L1	Cell division control protein 2 homolog	up	57	1bi7	NFKBIZ	Molecule possessing ankyrin repeats induced by lipopolysaccharid	down	57	
2	UHRF1	ubiquitin-like, containing PHD and RING finger domains, 1	up	54	1nbf	USP9Y	Probable ubiquitin carboxyl-terminal hydrolase FAF-Y	down	51	
3	RASAL2	RAS protein activator-like 2	up	70	1wq1	RAB27A	Ras-related protein Rab-27A	up	62	
4						RHOA	Transforming protein RhoA (H12)	up	72	
5						RAB2	Ras-related protein Rab-2A	up	51	
6						RAB22A	Ras-related protein Rab-22A	up	58	
7						RRAS	Ras-related protein R-Ras	up	89	
8						RAN	ras-related nuclear protein	up	68	
9						KRAS2	Transforming protein p21 (K-Ras 2)	up	100	
10						GEM	GTP-binding protein	up	51	
11						RASD1	Dexamethasone-induced Ras-related protein 1	down	62	
12						RERG	RAS-like, estrogen-regulated, growth inhibitor	down	68	
13	CTSG	Cathepsin G precursor	up	62	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
14				78	1ezx	SERPINI2	serine protease inhibito	down	58	
15	CDKN3	Cyclin-dependent kinase inhibitor 3	up	100	1fq1	CDC2L1	Cell division control protein 2 homolog	up	75	
16						DYRK2	Dual-specificity tyrosine-phosphorylation regulated kinase 2	up	58	
17						CDK7	Cell division protein kinase 7	up	70	×
18						CDC2	Cell division control protein 2	up	83	✓
19	MYL9	Myosin regulatory light chain 2	up	100	1dfk	MYH9	Myosin heavy chain	up	85	
20	MLRM	Myosin regulatory light chain 2	up	100	1dfk	MYH9	Myosin heavy chain	up	85	
21	KLK10	Kallikrein 10 precursor	up	62	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
22				71	1ezx	SERPINI2	Serine protease inhibitor	down	58	
23	BIN1	Myc box dependent interacting protein 1	up	100	1gri	GRB14	Growth factor receptor-bound protein 14	down	100	
24	TMPRSS4	Transmembrane protease, serine 4	up	81	1co7	TFPI2	Tissxue factor pathway inhibitor 2	down	78	
25				78	1ezx	SERPINI2	Serine protease inhibitor	down	58	
26				66	1sgf	NTF5	Neurotrophin-5 precursor	down	80	
27	CSTA	Cystatin A	up	70	1nb3	CTSC	Cathepsin C	up	100	
28				68	1nb3	CTSL	Cathepsin L	up	100	
29				73	1stf	CTSK	Cathepsin K	up	73	
30	ARHGDIA	Rho GDP-dissociation inhibitor 1 alpha	up	100	1cc0	RHOA	Transforming protein RhoA	up	100	✓
31						KRAS2	GTPase KRaT	up	52	
32	CHN1	N-chimaerin	up	57	1ow3	RHOA	Transforming protein RhoA (H12)	up	100	

	Protein 1	Description	up/ down	Interface conserved	Complex PDB ID	Protein 2	Description	up/ down	Interface conserved	Confirmed by literature
33				53	1tx4	ARL4A	ADP-ribosylation factor-like protein 4A	up	56	
34				53	1tx4	RAB22A	Ras-related protein Rab-22A	up	52	
35				59	1grn	KRAS2	Transforming protein p21 (K-Ras 2)	up	60	
36	PRSS3	Trypsin III precursor	down	85	1co7	TFPI2	Tissue factor pathway inhibitor 2 precursor	down	78	
37				92	1ezx	SERPINI2	Serpin I2 precursor	down	58	
38	CTRC	Caldecrin	down	75	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
39				85	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
40	ARHGAP6	Rho-GTPase-activating protein 6	down	63	1ow3	RHOA	Transforming protein RhoA	up	100	
41				62	1tx4	ARL4A	ADP-ribosylation factor-like protein 4A	up	56	
42				62	1tx4	RAB22A	Ras-related protein Rab-22A	up	52	
43				62	1grn	KRAS2	Transforming protein p21	up	60	
44	CETN2	Centrin 2	up	100	1dfl	MYH9	Myosin heavy chain, nonmuscle type A	up	87	
45				80	1m63	PPIG	Peptidyl-prolyl cis-trans isomerase G	up	80	
46				92	1m63	PPP3CA	Serine/threonine protein phosphatase 2B	up	53	
47				80	1m63	PPIF	Peptidyl-prolyl cis-trans isomerase, mitochondrial precursor	down	100	
48				80	1m63	PPID	40 kDa peptidyl-prolyl cis-trans isomerase	down	80	
49	CSTB	Cystatin B	up	66	1stf	CTSC	Dipeptidyl-peptidase I	up	94	
50				63	1stf	CTSL	Cathepsin L precursor	up	94	
51				73	1stf	CTSK	Cathepsin K precursor	up	94	
52	C2	Complement C2	up	55	1tfx	TFPI2	Tissue factor pathway inhibitor 2	down	53	
53				78	1ezx	SERPINI2	Serpin I2 precursor serine protease inhibitor)	down	58	
54				58	1sgf	NTF5	Neurotrophin-5 precursor	down	80	
55	KLK1	Kallikrein 1	down	71	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
56				71	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
57				66	1sgf	NTF5	Neurotrophin-5 precursor	down	80	
58				62	1tx4	KRAS2	Transforming protein p21 (K-Ras 2) (Ki-Ras) (c-K-ras)	up	56	
59	CDC2	Cell division control protein 2	up	57	1bi7	NFKBIZ	Molecule possessing ankyrin repeats induced by lipopolysaccharide	down	51	
60	PPIF	Peptidyl-prolyl cis-trans isomerase	down	100	1mf8	PPP3CA	Serine/threonine protein phosphatase 2B catalytic subunit, alpha isoform	up	77	
61	HPCAL1	Hippocalcin-like protein 1	down	60	1m63	PPIG	Peptidyl-prolyl cis-trans isomerase G	up	80	
62				100	1mf8	PPP3CA	Serine/threonine protein phosphatase 2B catalytic subunit, alpha isoform	up	55	

	Protein 1	Description	up/ down	Interface conserved	Complex PDB ID	Protein 2	Description	up/ down	Interface conserved	Confirmed by literature
63				60	1m63	PPIF	Peptidyl-prolyl cis-trans isomerase, mitochondrial precursor	down	100	
64				60	1m63	PPID	40 kDa peptidyl-prolyl cis-trans isomerase	down	80	
65	F11	Coagulation factor XI	down	75	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
66				71	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
67	F12	Coagulation factor XII	down	87	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
68				71	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
69				65	1sgf	NTF5	Neurotrophin-5 precursor	down	66	
70	KLKB1	Plasma kallikrein	down	75	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
71				64	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
72	CTRL	Proteasome subunit beta type 10	down	75	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
73				85	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
74	EL2B	Elastase 2A	down	78	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
75				92	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
76	HABP2	hyaluronan binding protein 2	down	75	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
77				78	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
78				65	1sgf	NTF5	Neurotrophin-5 precursor	down	66	
79	ELA3B	Elastase IIIB	down	65	1taw	TFPI2	Tissue factor pathway inhibitor 2	down	76	
80				64	1ezx	SERPINI2	Serpin I2 serine protease inhibitor	down	58	
81	DLC1	Rho-GTPase-activating protein 7	down	54	1ow3	RHOA	Transforming protein RhoA	up	100	
82				62	1tx4	ARL4A	ADP-ribosylation factor-like protein 4A	up	56	
83				62	1tx4	RAB22A	Ras-related protein Rab-22A	up	52	
84				69	1am4	RAN	ras-related nuclear protein	up	57	

Table 4.1: 84 predicted interactions where both partners are deregulated in the PDAC microarray experiment. *Protein 1* is the interaction partner of *Protein 2*. The complex column shows the PDB ID of the known complex template assigned by GTD. The interface conservation percentages of protein and their complex template are shown. Some of the predictions could be verified by checking the literature and are marked with ✓. The × sign represents a negative literature confirmation [Lolli et al., 2004].

4.1.3.4 Prediction of kinase-inhibitor interactions in upregulated pancreas tumour genes expression data

Kinases catalyse the transfer of a phosphate group from a donor, such as ADP or ATP to an acceptor. Cyclins combine with cyclin dependent kinases (CDKs) to form activated kinases that phosphorylate targets leading to cell cycle regulation. A breakdown in the regulation of this cycle can lead to out of control growth and contribute to tumour formation [Deshpande et al., 2005]. Defects in many of the molecules that regulate the cell cycle have been implicated in cancer. Moreover, protein kinases are elementary switches in signal transduction cascades and are overly important in the development of cancer as known from the activation of HER2/NEU in breast carcinoma [Menard et al., 2004]. Protein kinases are well investigated and a crucial target for anti-neoplastic therapy [Sawyer, 2004, Ishizawar and Parsons, 2004, Tibes et al., 2005]. Therefore the potential regulation of these kinases in pancreatic cancer is important to further understand this disease. For this example, we used the data set described in case study I where we only consider the genes which are over expressed.

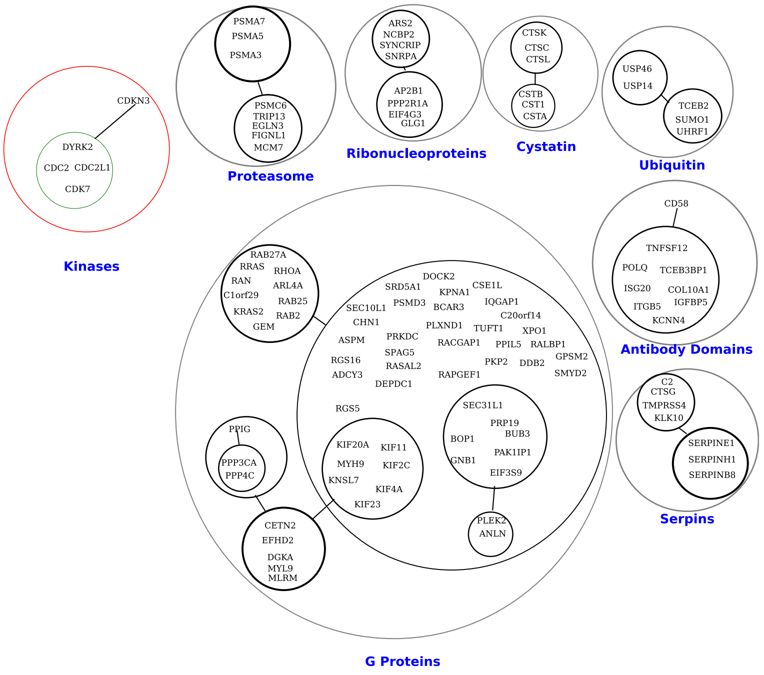
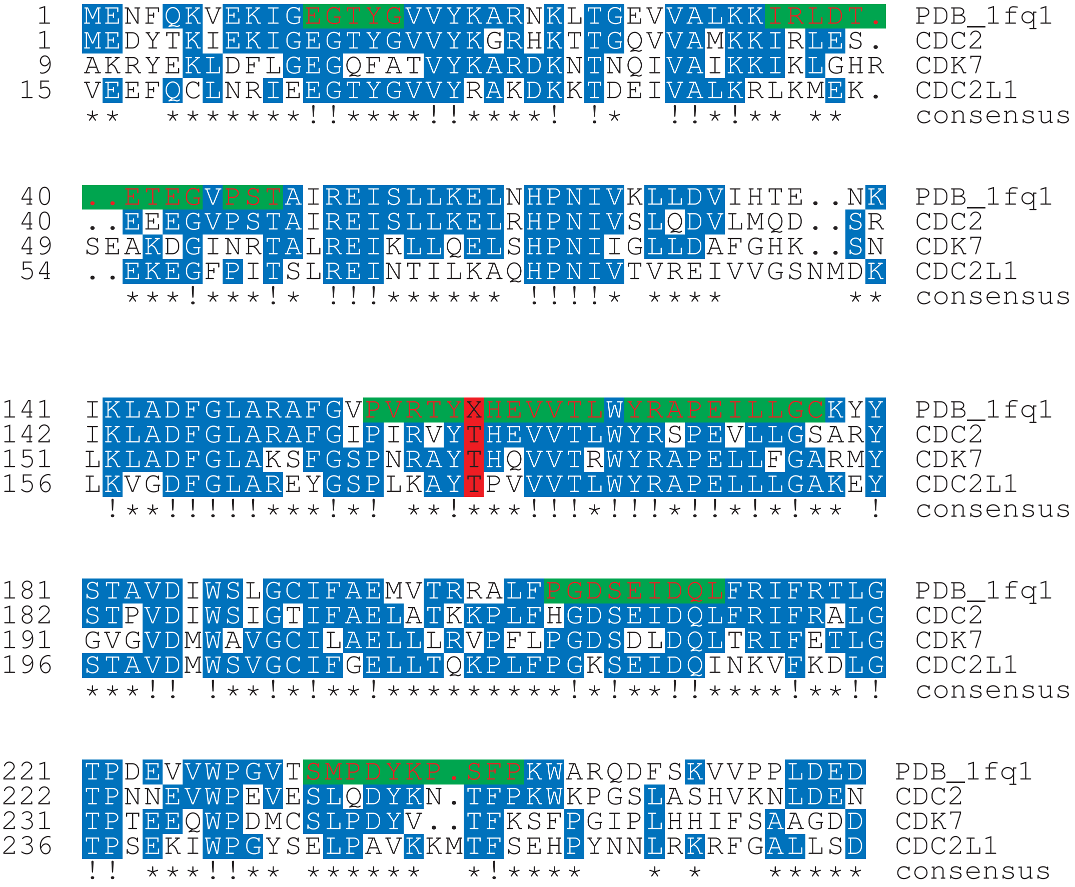


Figure 4.5: For the overexpressed genes in the pancreas data set there are eight distinct clusters of interactions. The clusters can be broadly classified as kinase/inhibitor, G proteins, DNA replication/proteasome subunits, ubiquitin, cystatin, serine proteases, ribonucleoproteins and antibody domains.

Figure 4.5 shows the resulting eight different interaction subnetworks. Within the set of upregulated genes, each of the eight subnetworks identifies a group of genes from different

Protein	SID	e-value	Interface SeqId	Thr160 conserved
CDC2	d2phka0	e-108	75.4%	√
CDK7	d2phka0	5e-69	52.6%	√
CDC2L1	d2phka0	6e-70	57.9%	√



```
  1  MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDT.  PDB_1fq1
  1  MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLES.  CDC2
  9  AKRYEKLDFLGEGQFATVYKARDKNTNQIVAIKKIKLGHR  CDK7
 15  VEEFQCLNRIEEGTYGVVYRAKDKKTDEIVALKRLKMEK.  CDC2L1
     **  *******!!****!!****! !*  !!*!** **   consensus

 40  ..ETEGVPSTAIREISLLKELNHPNIVKLLDVIHTE..NK  PDB_1fq1
 40  ..EEEGVPSTAIREISLLKELRHPNIVSLQDVLMQD..SR  CDC2
 49  SEAKDGINRTALREIKLLQELSHPNIIGLLDAFGHK..SN  CDK7
 54  ..EKEGFPITSLREINTILKAQHPNIVTVREIVVGSNMDK  CDC2L1
       ***!***!* !!!****** !!!!* ****      **  consensus

141  IKLADFGLARAFGVPVRTYXHEVVTLWYRAPEILLGCKYY  PDB_1fq1
142  IKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSARY  CDC2
151  LKLADFGLAKSFGSPNRAYTHQVVTRWYRAPELLFGARMY  CDK7
156  LKVGDFGLAREYGSPLKAYTPVVVTLWYRAPELLLGAKEY  CDC2L1
      !**!!!!!***!*! **!***!!!*!!!*!!*!*!** !  consensus

181  STAVDIWSLGCIFAEMVTRRALFPGDSEIDQLFRIFRTLG  PDB_1fq1
182  STPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALG  CDC2
191  GVGVDMWAVGCILAELLLRVPFLPGDSDLDQLTRIFETLG  CDK7
196  STAVDMWSVGCIFGELLTQKPLFPGKSEIDQINKVFKDLG  CDC2L1
     ***!! !**!*!**!*********!*!**!!****!**!!  consensus

221  TPDEVVWPGVTSMPDYKP.SFPKWARQDFSKVVPPLDED  PDB_1fq1
222  TPNNEVWPEVESLQDYKN.TFPKWKPGSLASHVKNLDEN  CDC2
231  TPTEEQWPDMCSLPDYV..TFKSFPGIPLHHIFSAAGDD  CDK7
236  TPSEKIWPGYSELPAVKKMTFSEHPYNNLRKRFGALLSD  CDC2L1
     !! ***!!** ******  *!****   * *   *****  consensus
```

functional categories. Out of the eight subnetworks, the kinase cluster is considered in more detail.

The genes interaction with CDKN3 are listed in Table 4.1 and are identified as kinases, and since there is an interaction of a kinase with the kinase inhibitor CDKN3, all of these kinases potentially interact with this inhibitor. As for all the genes in Table 4.1 the CDKN3 interaction with the four kinases have high confidence GTD assignments and the assigned structures align structurally well with the kinase with PDB structure 1fq1, chain b (all RMSDs are below 2 $\AA$).

One of these interactions is verified in literature [Hannon et al., 1994], namely CDC2 and CDKN3 and in the interaction databases DIP [Xenarios et al., 2000] and BIND [Bader and Hogue, 2000]. CDKN3 has been shown to interact with, and dephosphorylate the cyclin-dependent kinase CDK2 preventing its activation [Poon RY, 1995]. This gene was reported to

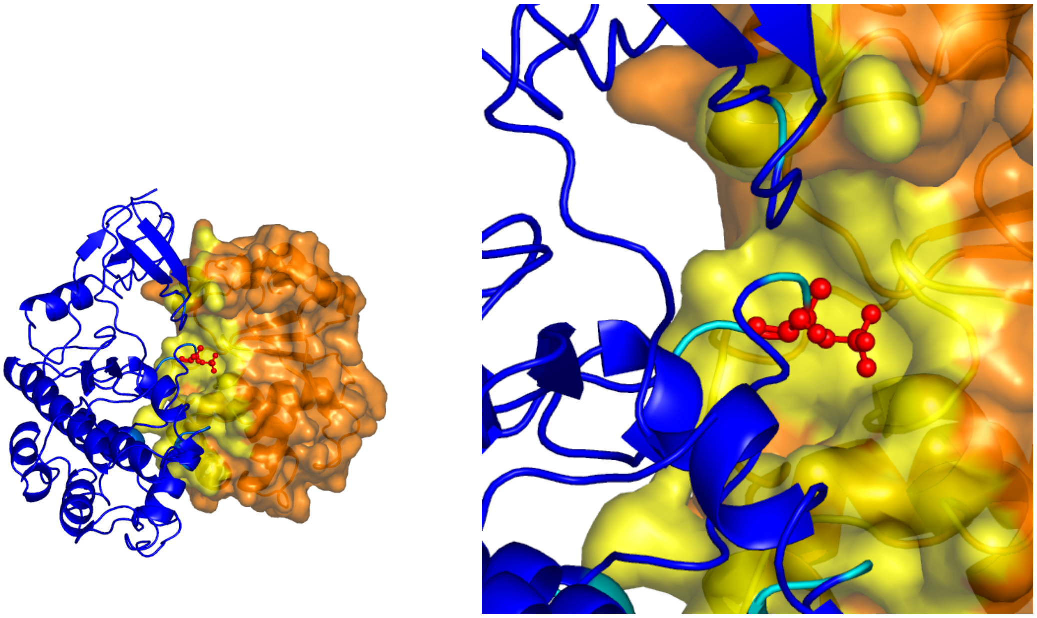


Figure 4.6: **Left**: Cyclin-dependent kinase 2 CDK2 (blue) interacting with the cyclin-dependent kinase inhibitor CDKN3 (yellow/orange). The interfaces are displayed in light blue and yellow, respectively. The phosphorylated threonine of CDK2, which protrudes into a pocket of the inhibitor, is shown in red balls-and-sticks mode. PDB ID 1fq1. **Right**: A closeup showing the phosphorylated threonine of CDK2, which protrudes into a pocket of the inhibitor

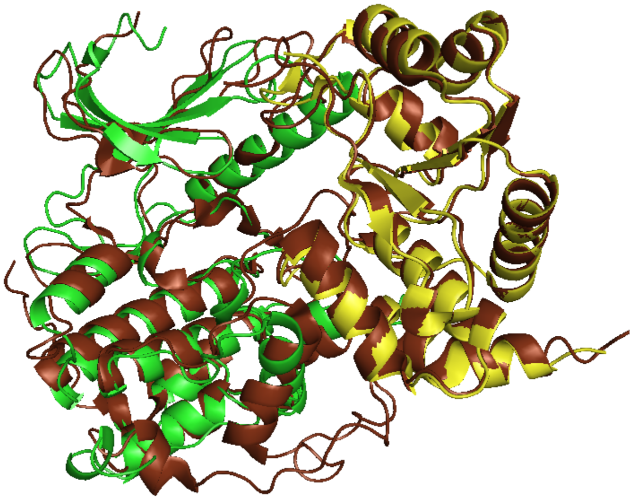


Figure 4.7: Structural alignment of the kinases CDK2 (PDB ID 1fq1, chain b) and PDB ID 2phk with alignment RMSD of 1.54929. The inhibitor (PDB ID 1fq1, chain a) is aligned to PDB ID 1fpz, chain a with alignment RMSD of 0.84583. 2phk and 1fpz are the structures assigned by GTD to CDK2 and CDKN3 and 1fq1 is the structure that shows the interaction of the two domains.

be deleted, mutated, or overexpressed in several kinds of cancers. To validate the interactions we considered the sequence alignments of CDC2's structure (PDB ID 1fq1) with the other kinases. We found only for CDK7 and CDC2L1 ($> 50\%$) sequence identity with the threaded structure and only in these kinases the aligned interface residues are well conserved ($> 50\%$). In particular, the key residue threonine 160 (see also Fig. 4.6) is conserved.

An example of a false positive interaction In [Lolli et al., 2004], the authors experimentally confirmed that CDK7 is not a substrate for CDKN3, even though CDK7 was found to be interacting with CDKN3 according to our method.

CDK7 is known to be important regulator of cell cycle progression. This protein is thought to serve as a direct link between the regulation of transcription and the cell cycle. The activation segment is phosphorylated at Thr170 and is in a defined conformation that differs from that in phospho-CDK2 and phospho-CDK2/cyclin A [Lolli et al., 2004]. Its expression and activity are constant throughout the cell cycle.

Grasping the molecular basis of specificity which is defined by the mechanism of how proteins discriminate their natural binding partners from many other possible ligands with similar sequences and structures is a major challenge for prediction of protein-protein interactions. This example sheds light on the challenges of protein-protein interaction specificity, specially of interactions between kinases and cyclins.

4.1.4 Validation of candidates.

4.1.4.1 Docking

For the last two decades docking has been successfully used in drug discovery where for example cases like how drug and enzyme or receptor of protein, fit together or how the action of a harmful protein in human body may be prohibited by finding an inhibitor, which binds to that particular protein. Molecular docking process involves the prediction of the correct relative orientation of two or more molecular structures when bound to each other to form a stable complex. The main goals of docking studies are finding an accurate structural models and the correct prediction of biological activity. Most classic docking techniques aim to find the best docked complex by using scoring functions that are evaluated on the basis of calculations of approximate shape and electrostatic complementarity. Docking usually requires knowledge of where the binding sites are. This knowledge is used to point out preferred orientation which in turn can be used to predict the strength of binding affinity between two molecules [Aloy and Russell, 2006]. We used BDOCK [Huang and Schroeder, 2005] for docking the TMPRSS4–TFPI2 complex.

4.1.4.2 Homology modelling

Homology modelling is a a class of methods of protein-structure prediction that identify or more known protein structures likely to resemble a known structure as a modelling template for a homologue that has been identified on the basis of sequence similarity [Aloy and Russell, 2006]. We used MODELLER version 8v0 [Mart-Renom et al., 2000] to confirm that the GTD fold prediction is similar to the interaction template.

4.1.4.3 Molecular Dynamic

For the Molecular Dynamics experiments conjugate gradient energy minimisation using NAMD [Phillips et al., 2005] with the CHARMM22 force field were applied. For the simulation on the TMPRSS4–TFPI2 complex, we observed a stabilisation of the complex after 10,000 steps.

Molecular Dynamics simulations confirmed that the predicted TMPRSS4–TFPI2 interaction remains stable. The Molecular Dynamics experiments were performed by Anne Tuukkanen, Biotec Dresden.

4.1.5 Summary and Conclusions

In this study, we use the integrative approach described in chapter 3 to identify key interactions and pathways from a set of genes. We apply this approach to a data set of genes deregulated in pancreatic cancer. As a first step, we construct a pathway network from the deregulated cancer genes. The analysis of such a network gives an overview to explain how the pathways affect each other, resulting in tumorigenesis. In the case of PDAC, we find most pathways previously reported to be involved in cancer. These include signal transduction, immune system, cell growth and death, signalling molecules and interaction, cell motility, and cell communication.

In addition, we observe the alteration of the calcium pathway. We conclude that it plays an important role in pancreas specific tumorigenesis.

The method builds on a number of structural data sources such as PDB, SCOP, GTD, and SCOPPI. We apply the method to our data set of deregulated pancreatic cancer genes. As a result, we predict 81 novel interactions that are specific for the underlying disease. We map these interactions onto a well-known picture of cancer hallmarks and draw a network of all predicted interactions as well as literature confirmed interactions. We observe that most of the literature confirmed interactions are located inside the cell, whereas the predicted interactions are mainly taking place between transmembrane and extracellular proteins. One reason for this bias could be that transmembrane proteins are more difficult to study experimentally than cytosolic proteins. The interactions found may prove valuable to improve our understanding of the regulatory mechanisms underlying the development of pancreatic cancer. Finally, we examine two examples in detail. The first example is the predicted interaction between TMPRSS4 and TFPI2. We believe that TFPI2 naturally inhibits the TMPRSS4 protease. Since we find TFPI2 to be downregulated in pancreatic cancer, TMPRSS4 might be able to facilitate tissue invasion and metastasis. Another indication for the importance of the role of this interaction as potential drug target, is the patent [Park et al., 2007] of an anticancer drug comprising inhibitor of TMPRSS4. The authors claim that it can be used effectively for the treatment of cancer by inhibiting TMPRSS4 expression in cancer cells and thereby inhibiting cancer cell invasion and cancer cell growth.

The second example is another interesting predicted interaction between the kinase inhibitor CDKN3 and CDC2L1. For this analysis we only considered the over expressed genes from the dataset. We identify eight interaction networks, and we considered a kinase-inhibitor cluster in detail. This analysis reveals an interaction between CDC2 and the inhibitor CDKN3, which is documented in the literature and a novel interaction CDC2L1, which we believe to be valid as over 50% of the interaction interface is conserved and a key residue is also conserved. The interactions may prove valuable to improve our understanding of the regulatory mechanisms underlying the development of pancreatic cancer.

4.2 Case study II: Potential therapeutic targets for the treatment of PDAC using a novel drug (BVDU)

4.2.1 Introduction

Treatment of pancreatic cancer
New treatment strategies for resectable and unresectable pancreatic cancer are under active investigation. One of the standard and most widely used drug for treatment of pancreatic cancer is Gemcitabine-based chemotherapy. Recently, these standard chemotherapies were found to give better results when combined with specific substances sensitising the tumour towards chemotherapy. These treatments include, combinations of Gemcitabine with other adjuvant therapies [Khosravi and Daz, 2005]. Standard chemotherapy drugs are usually a DNA damaging substance that replaces cytidine which is one of the building blocks of nucleic acids, during DNA replication. The process arrests tumour growth, as new nucleotides cannot be attached to the ”faulty” nucleoside, resulting in apoptosis. After prolonged therapy with Gemcitabine the cancer cells acquire resistance to the drug which critically limits the outcome of the treatment [Gennatas et al., 2006].

4.2.2 Chemoresistance

Frequent chemotherapeutic treatment induces chemoresistance of remaining cancer cells by altering gene expression and inducing genomic instability because of mutations, recombination, and gene amplification events. The molecular mechanisms that contribute to the resistance of PDAC to various anticancer therapies are not well understood. Chemoresistence mainly involve two major pathways, programmed cell death (apoptosis) and survival pathways. Factors like RNA interference targeting focal adhesion kinase can also enhance pancreatic adenocarcinoma Gemcitabine chemosensitivity [Duxbury et al., 2003]. The loss of BNIP3 expression is a late event in pancreatic cancer contributing to chemoresistance and worsened prognosis [Erkan et al., 2005]. Intrinsic and acquired resistance to chemotherapy critically limits the outcome of cancer treatments. For many years, it was assumed that the interaction of a drug with its molecular target would yield a lethal lesion, and that determinants of intrinsic drug resistance should therefore be sought either at the target level (quantitative changes or/and mutations) or upstream of this interaction, in drug metabolism or drug transport mechanisms. It is now apparent that independent of the factors above, cellular responses to a molecular lesion can determine the outcome of therapy [Pommier et al., 2004].

During the implementation of a long term screening program for inhibitors of chemoresistance, [(E)-5-(2-bromovinyl)-2’-deoxyuridine (BVDU) was the only identified substance of clinical relevance [Fahrig et al., 2003]. The effect of BVDU, which supports apoptosis and prevents the acquisition of chemoresistance, was demonstrated *in vitro* and in patients with pancreatic cancer [Fahrig et al., 2006]. BVDU co-treatment significantly enhanced survival and time to progression [Fahrig et al., 2006]. These results encouraged the authors of [Fahrig et al., 2006] to investigate the effect of BVDU on pancreatic cancer patients. The resulting data set from this investigation was the starting point to apply our network approach to analyse the data to provide a comprehensive over view of how the genes involved act together to achieve such behaviour.

Material and methods

4.2.2.1 Dataset

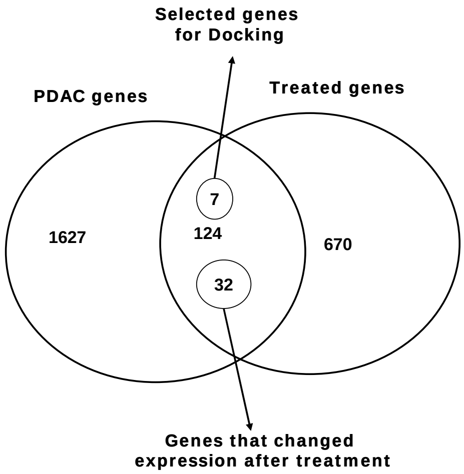

Selected genes for Docking
PDAC genes
Treated genes
7
1627
124
670
32
Genes that changed expression after treatment

Figure 4.8: The dataset consists of the gene expression levels of 670 genes. We compare this data set to the PDAC tumour microarray data set used in case study I which consists of 1627 genes. The overlap between the two data sets results in a new list of 124 genes. We further investigated the 32 which changed their expression to the opposite after treatment with the combination of Gemcitabine/BVDU

The dataset consists of genes obtained from pancreas carcinoma cell lines and pancreatic cancer patients. The expression of those genes was monitored before and after being treated by the combination of Gemcitabine/BVDU versus BVDU, Gemcitabine/BVDU versus Gemcitabine, Gemcitabine/BVDU versus control, BVDU versus control and Gemcitabine versus control. The dataset consists of the gene expression levels of 670 genes out of which 537 genes were treated with the combination of Gemcitabine/BVDU. We compare this data set to the the PDAC tumour microarray data set used in case study I which consists of 1627 genes. The dataset is obtained by integrating various analyses of the gene expression profile of PDAC from Affymetrix GeneChip experiments and the meta-analysis of PDAC gene expression profiles from public available data of other projects [Grutzmann et al., 2005].

4.2.2.2 Expression altering after treatment

The overlap between the two data sets results in a new list of 124 genes. We further investigated the 32 genes described in Table 4.3 which changed their expression to the opposite after treatment with the combination of Gemcitabine/BVDU (i.e. the expression of KLKB1 changed

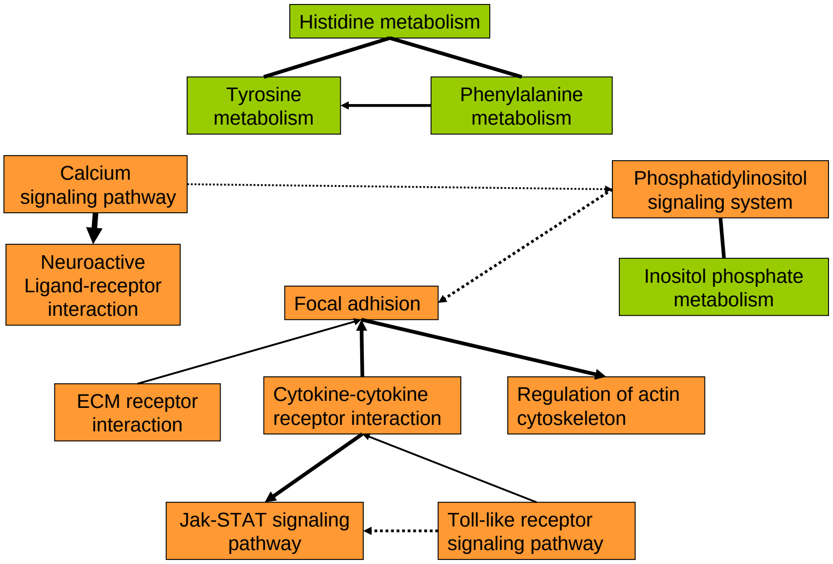

Histidine metabolism
Tyrosine metabolism
Phenylalanine metabolism
Calcium signaling pathway
Phosphatidylinositol signaling system
Neuroactive Ligand-receptor interaction
Inositol phosphate metabolism
Focal adhision
ECM receptor interaction
Cytokine-cytokine receptor interaction
Regulation of actin cytoskeleton
Jak-STAT signaling pathway
Toll-like receptor signaling pathway

Figure 4.9: A network of the highly involved pathways in the dataset. Pathways are connected by a solid line if they share at least 4 genes. Pathways are coloured differently (orange for signalling pathways and green of metabolic pathways). Pathways that are not sharing genes but are down or upstream of each other (according to KEGG) are represented by a doted line. We observe an overerepresentation of signalling and cell communication pathways which can explain the mechanism of the cancer recruiting the focal adhesion, cytokine-cytokine receptor interaction to pass the signals to the ECM to activate the metastasis process that define the PDAC.

from upregulated in the PDAC data set to downregulated after treatment). This set of cancer genes are of greater interest to us because they respond to the BVDU drug by altering their expression after treatment.

4.2.3 Data analysis approach

The biological roles of most proteins, are characterised by which other macromolecules they interact with. We applied our structural template method to predict potential interactions among the genes of the data set. The same evaluation criterion was applied. An interaction network consisting of predicted and known interaction in constructed. Protein interaction databases that uses data mining technique were queried to identify known interactions. We use the Gene ontology to annotate the network. The genes of our data set are assigned to KEGG pathways, then a network with pathways that are connected (if sharing 4 or more genes) is constructed.

4.2.4 Detailed study

We further investigated seven genes out of the 32. The choice of these seven genes was based on an intensive literature search to select those known to be oncogenes or related to pancreatic cancer or novel genes with no information about them. Six of these genes are of known structures, the structure of Interferon induced protein 44 like (IFI44L) was obtained using threading see Table 4.4.

Docking In order to gain insight of pathways mechanisms and their interference with drugs, in particular BVDU, which is known to have an effect on the seven selected oncogenes. We perform docking experiments of the BVDU molecule with these seven genes to evaluate out hypothesis. We used PatchDock, a molecular Docking Algorithm Based on Shape Complementarity Principles [Schneidman-Duhovny et al., 2005] for docking all seven genes with the BVDU molecule. The PatchDock algorithm works as follows, it first computes a molecular shape representation of surface the molecules, then geometric patches (concave, convex and flat surface pieces) are identifies, these patches are then filtered and only patches which are identified as hot spots are retained. Concave patches are matched with convex and flat patches with any type of patches. For filtering and scoring they discard all complexes with unacceptable penetrations of the atoms of the receptor to the atoms of the ligand. Finally, the remaining candidates are ranked according to a geometric shape complementarity score. The docking solutions of the Nicotinamide N-methyltransferase (NNMT), kinesin family member 20A (KIF20A) and Transmembrane serine protease 4 (TMPRSS4) protein structures with the BVDU molecule are shown in Figure 4.10, 4.11 and 4.12 respectively.

4.2.5 Results and Discussion

We performed the pathways analysis method on all the dataset genes. The result of the analysis is a network of the highly involved pathways in the dataset. Figure 4.9 shows the pathway network. Pathways are connected by a solid line if they share at least 4 genes. Pathways are coloured differently (orange for signalling pathways and green of metabolic pathways). Pathways that are not sharing genes but are down or upstream of each other (according to KEGG) are represented by a doted line. We observe an over-representation of signalling and cell communication pathways which can explain the mechanism of the cancer recruiting the focal adhesion, cytokine-cytokine receptor interaction to pass the signals to the ECM to activate the metastasis process that define the PDAC. BVDU is known to enhance survival time in patients with pancreatic cancer.

4.2.5.1 A proposed mechanism of action for BVDU.

We further performed docking experiments which indicate that BVDU is able to bind to the active site of TMPRSS4, KIF20A and NNMT. The docking results indicate that the three proteins are potential targets of BVDU.

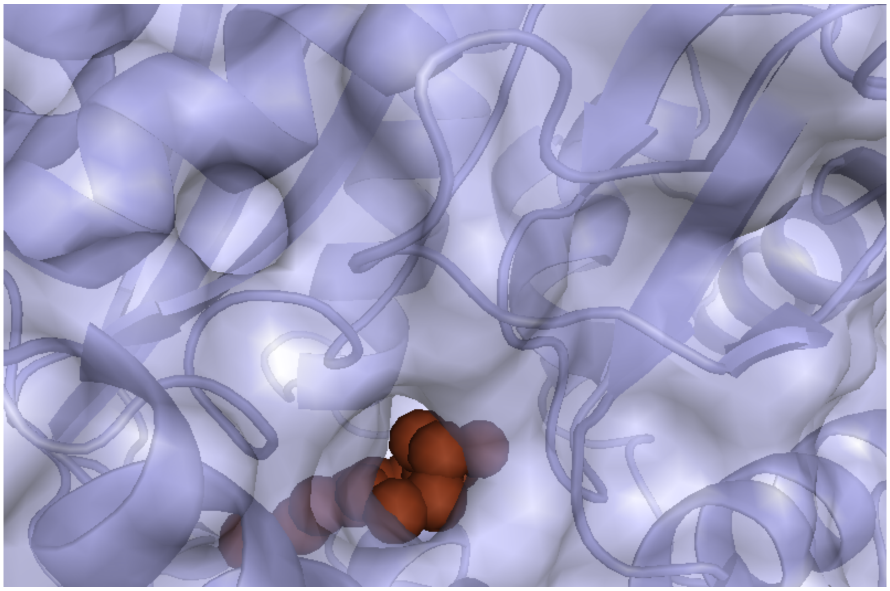


Figure 4.10: Docking of the protein Nicotinamide N-methyltransferase (NNMT) (light blue) with BVDU (brown). Its function is to catalyse the N-methylation of nicotinamide and other pyridines to form pyridinium ions. This activity is important for biotransformation of many drugs and xenobiotic compounds protein.

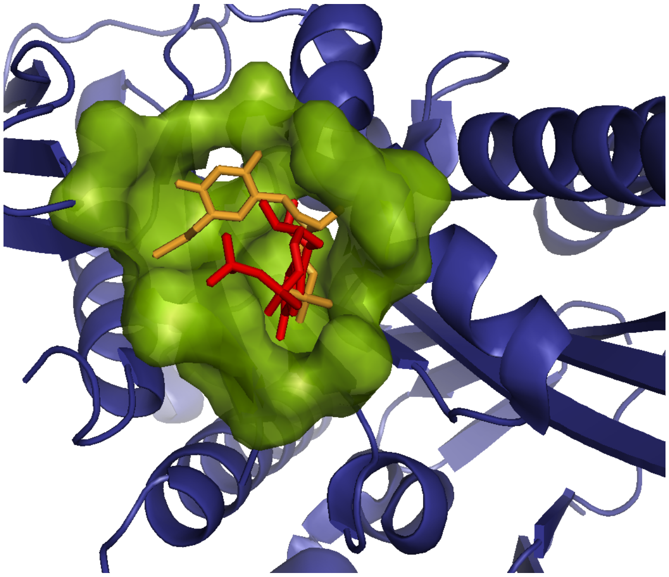


Figure 4.11: The docking of protein structure Kinesin family member 20 A (KIF20A) (blue) with BVDU (orange) which is responsible for the transport of Golgi membrane. The KIF20A protein has two ADP binding sites the interface of one of them is shown in green. BVDU docked to the same pocket as the ADP (red) probably mimicking its function.

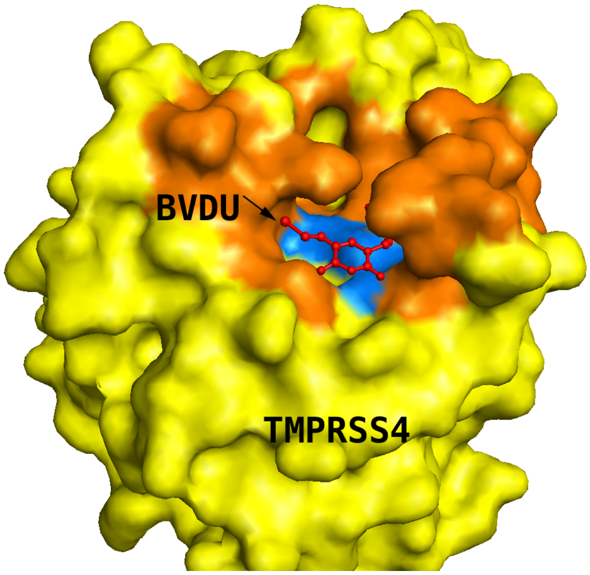


Figure 4.12: Docking of BVDU instead of its natural inhibitor suggests a potential interaction and role of BVDU as a TMPRSS4 inhibitor.

TMPRSS4: The relationship between metastasis and chemoresistance might indicate that acquired resistance to apoptosis as a result of chemotherapy could favour the metastatic process [Mehlen and Puisieux, 2006]. Since TMPRSS4 is known to play a role in tissue invasion and metastasis, we used PatchDock [Schneidman-Duhovny et al., 2003] to dock BVDU to TMPRSS4. The result is shown in Figure 4.12. BVDU clearly blocks the pocket with the active site. We can only speculate about the affinity of BVDU towards TMPRSS4 and that it could act as a competitive inhibitor for the natural substrate of TMPRSS4.

KIF20A: Taniuchi et al. reported that the down-regulation of KIF20A, a kinesin involved with membrane trafficking of discs large homologue 5, can attenuate growth of pancreatic cancer cell. This can explain another action mechanism for BVDU since docking results in Figure 4.11 suggests that the BVDU downregulates the KIF20A protein causing the slow down of the growth and metastasis of the cancer as confirmed by [Taniuchi et al., 2005].

NNMT: NNMT function is to catalyses the N-methylation of nicotinamide and other pyridines to form pyridinium ions. This activity is important for biotransformation of many drugs and xenobiotic compounds. The docking result of BVDU with the NNMT structure suggests that the BVDU binds to the binding sites of NNMT see Figure 4.10. The binding site data is obtained from the swissprot database.

Limitations of small molecule docking PatchDock is a geometric docking method that apply fast geometric scoring and search and avoids exhaustive orientaion search. In comparison to other docking methods it is fast and easy to use for non docking experts who are mainly interested in fast general results. We are aware of the fact that using small molecules such as BVDU for docking experiments can produce lots of false positive results. Ongoing experiments by Farbig et. al suggests that the biologically relevant binding partner of BVDU is not among the three above mentioned examples (results not published). Docking in this case could work as initial filter to remove candidates that are not likely to bind at all to the molecule in hand.

4.3 Case study III: Protein-protein interaction highlights the importance of the co-expression of KLK6 and KLK10 as prognostic factor for survival in pancreatic ductal adenocarcinoma

4.3.1 Introduction

For case study III and IV, all the wet lab experiments were conducted by Rückert and colleagues from the University Hospital Carl Gustav Carus, Dresden.

The second case study was conducted to find and analyse prognostic factors for the survival of PDAC patients. Pancreatic carcinoma shows an unsatisfactory response to oncological treatment. This demonstrates the need for new therapeutic approaches and also for biomarkers, which make early diagnosis possible. Recently, Grützmann et al., Iacobuzio-Donahue et al. and Yousef et al. have shown that human kallikrein 10 and human kallikrein 6 are among the most highly and specifically overexpressed genes in pancreatic cancer compared to normal and benign pancreas tissues. KLK10 and KLK6 are members of the kallikrein family of 15 known proteases in humans, which play an emerging role in tumour micro-environment, invasion and angiogenesis [Borgono and Diamandis, 2004]. Kallikreins exert this function as secreted trypsin and chymotrypsin-like proteases by degradation of the extracellular matrix, which is an important reservoir for cytokines and growth factors [Borgono and Diamandis, 2004]. This highlights the importance of kallikreins as candidate genes for diagnosis and therapy in pancreatic cancer. Therefore, the aim of this study was to evaluate the role of these kallikreins and their value as biomarkers in PDAC using protein-protein interactions and experimental methods The function of the KLK10 protein is poorly documented, neither the activators nor the substrates for KLK10 are actually known [Zhang et al., 2006]. KLK6, is highly expressed in several malignancies like ovarian, breast, colon or gastric cancer [Nagahara et al., 2005, Yousef et al., 2004]. It is correlated with lymphatic invasion and poor prognosis in gastric cancer. KLK6 might exert this effect by degradation of matrix proteins and thereby augmentation of cancer cell motility and proliferation [Ghosh et al., 2004]. In this study, Rückert et al. could show experimentally that kallikrein 10 and 6 demonstrate a strong protein expression in pancreatic carcinoma and are associated with poor patient prognosis and thereby might contribute to the aggressive character of this malignancy. The results indicate that this effect is most likely mediated by interaction of KLK6 with factors of the extracellular matrix and enhancement of cancer cell motility by KLK10.

4.3.2 Data set

We use the data set described in case study I where we only consider KLK10 and KLK6 and their known and potential interacting partners.

4.3.3 Material and methods

To find possible novel interactions we applied our protein-protein interaction prediction methods as previously described in Chapter 3 for KLK6 and KLK10 which we hypothesise that it can explain poor survival of PDAC patients. We also queried databases with known protein-protein interactions such as NetPro, BIND and HPRD. A number of experimental techniques were used to test the relation between the co-expression of KLK10 and KLK6 and their function as prognostic factors for survival in PDAC.

4.3.3.1 Brief summary of the experimental methods

Immunohistochemistry study and evaluation Immunohistochemistry is the process of localising antigens or proteins in tissue sections exploiting the principle of antibodies binding specifically to antigens in biological tissues by the use of labelled antibodies as specific reagents through antigen-antibody interactions that are visualised by a marker. For the experimental details see [Rückert et al., 2008].

Construction of a virtual subarray and bioinformatics analysis For the construction of the virtual subarray, the experimental group used data obtained from an Affymetrix GeneChip using extracted RNA from microdissected tissue as described earlier by Pilarsky and colleagues. Genes were scored as differentially expressed if they displayed a fold change > 2. To identify signature genes the method described in [Grutzmann et al., 2004b] was used.

4.3.4 Results and Discussion

KLK10 and KLK6 are among the most highly and specifically overexpressed genes in pancreatic cancer compared with normal and benign pancreas tissues [Grützmann et al., 2003, Iacobuzio-Donahue et al., 2003a, Yousef et al., 2004]. This study confirmed a marked overexpression of KLK10 in PDAC by means of a virtual subarray. Immunohistochemistry in native tumour tissue could prove not only an intense expression for KLK10 in 64.4% of the malignant cells, but also for KLK6 in 91.5%. Both proteins were located in the cytoplasm, from where they are likely to be secreted [Borgono and Diamandis, 2004]. Co-expression of different kallikreins, similar to the situation found in our study, was already reported in skin and different glands. In these tissues the kallikreins can act independently, but also together as part of proteolytic cascades [Petraki et al., 2002 Jun-Jul, Borgono and Diamandis, 2004]. The latter seems to be an important mechanism in pancreatic cancer, because expression of KLK10 itself could not be associated with poor survival in PDAC, whereas the co-expression of both kallikreins was significantly associated with poor survival.

It is most interesting, in which ways kallikreins affect cellular signalling and thereby contribute to cancer progression. It was already reported that KLK6 is known to influence communication between malignant cells and their environment by degradation of extracellular matrix and thereby facilitate tumour invasion and metastasis [Borgono and Diamandis, 2004] In contrast, functional data on KLK10 are very limited. Although Zhang et al [Zhang et al., 2006] suggested, that KLK10 was not even an active protease, it was stated in the same report that neither the protein relevant for conversion of KLK10 into its active form nor the physiological substrates for KLK10 are known. So, the importance of KLK10 in tumour progression remains unclear.

It therefore seems crucial to further pinpoint some of the components, which might be responsible for the pathophysiological effect of KLK10. To find possible interaction partners for both kallikreins we used the our method of protein interaction prediction.

These potential interaction partners of KLK10 are Tissue factor pathway inhibitor 2 (TFPI2) and Protease inhibitor (SERPINI2). These interactions implies that it might have a role in the pathophysiology of PDAC see Table 4.5. The protein-protein interaction databases NetPro, HPRD and BIND provided information about the known interaction partners of KLK6 which included alpha1 antiproteinase (SERPINF2), which an inhibitor for KLK6 action, the interaction with Antithrombin-III precursor AT III (SERPINC1) shows a branching between kallikreins and blood coagulation cascade, as already previously reported. Another interaction partner we found is synuclein, which integrates presynaptic signalling and membrane trafficking in neurons. The high expression of KLK6 might thereby play an important role in various pathologic processes of pancreatic cancer.

In conclusion, this study shows that KLK10 and KLK6 co-expression has an unfavourable influence on the survival in patients with PDAC. This effect might be mediated by direct or indirect interaction of the two kallikreins. The pathophysiological mechanisms are most likely degradation of the extracellular matrix and interaction with angiogenic factors by KLK6, whereas KLK10 augments cell motility. However, our findings suggest a high complexity of interactions between the kallikreins, which leaves it difficult to generally make statements about properties of single kallikreins.

The potential interaction of the upregulated KLK10 with the two downregulated inhibitors TFPI2 and SERPINI2 is a hypothesis that should be validated in vivo. Consequently, it might be possible to use inhibitors of kallikreins to disrupt interactions between the tumour and its environment and thereby reduce disease progression in patients with pancreatic cancer.

4.4 Case study IV: Pancreatic cancer related apoptosis pathway

4.4.1 Introduction

Resection of PDAC tumour is the only treatment with curative intent. It is only possible in about 10% of the patients because of the late clinical manifestation and the unfavourable location of the malignancy. Although more than 30 anti-cancer drugs have been described, the result of treatment with these alone or in combination with radiation are unsatisfactory. Given this dismal prognosis of PDAC, virtually every therapeutic class of anticancerous agents has been or is being investigated in advanced stages of this disease, but only with unsatisfying results [Ducreux et al., 2004]. Apoptosis resistance of pancreatic cancer cells seem to play an important part in this treatment failure, because most current cancer therapies for solid tumours like chemotherapy, ionising radiation and immunotherapy exert their anti-tumour effect on cells by inducing cell death in terms of apoptosis [Brown and Attardi, 2005, Schulze-Bergkamen and Krammer, 2004]. Recent publications highlighted the importance of different defects in the apoptosis pathway for the resistant behaviour of pancreatic cancer [Jones et al., 2008, Hamacher et al., 2008, Trauzold et al., 2003, Gukovskaya and Pandol, 2004, Westphal and Kalthoff, 2003].

Considering that there seem to be multiple defects in the apoptotic pathway, Rückert et. al. hypothesised that rather than seeking to target individual genes, agents that broadly target key nodal points may be preferable to overcome apoptosis resistance in pancreatic cancer. The aim of this study was therefore to identify key nodal points and to introduce a new method for simultaneous silencing of different target genes in PDAC. As a first step, we visualised the complex interactions of the cell death pathway by construction of a comprehensive map of the apoptosis signal transduction, which we validated computationally. The map contained 100 genes that represent the common cell death-signalling pathway which were manually picked by experts of the fields using literature search. The map was evaluated and validated by computational analysis techniques using interaction databases and protein-protein interaction prediction using inference from known interacting structural templates. The method which is described in chapter 3 rendered novel protein-protein interactions, three of which are discussed in detail in this chapter. By literature search and DNA-microarray analysis we could identify several candidate target genes. Especially defects at the level of the mitochondrial pathway might have a clinical relevance because the mitochondrion amplifies signals mediated by cell death receptors and additionally initiates the effects of radio- and chemotherapies [Fulda and Debatin, 2006]. Therefore, upregulated members of the intrinsic pathway, namely Bcl-2, XIAP and Survivin were chosen for simultaneous gene silencing.

Based on our results we conclude that XIAP, Survivin and Bcl-2 may play a role in inhibiting the intrinsic pathway in pancreatic cancer cells. Further, simultaneous gene silencing seems a beneficial method to increase the effect of gene silencing in contrast to single gene silencing.

The aim of this study is to construct a comprehensive dynamic mapping of the apoptosis pathway by integrating gene expression data of apoptosis-associated genes, whose topology is based on known and novel (predicted) molecular interactions. This map will be a useful source for studying changes and effects of different experiments on the apoptosis pathway such as gene silencing and cancer drugs mechanism and their side effects.

4.4.2 Material and Methods

4.4.2.1 Analysis of the literature

The apoptosis pathway related genes were assembled from electronic databases, such as the KEGG database, Gene Data Base of the National Center for Biotechnology Information[1] and GeneMAPP[2]. These data were manually supplemented and completed with genes identified in

[1] www.ncbi.nlm.nih.gov

[2] www.genmapp.org

publications. The literature search comprised publications until April 2008.

4.4.2.2 Computational validation of the apoptosis pathway

To validate the interactions depicted as arrows in the constructed apoptosis maps pathway (see Figure. 4.13), we queried databases with known protein-protein interactions such as NetPro[3], SCOPPI[4] and HPRD[5] and compared them to our data. Furthermore, to find novel interactions we applied the structure-based prediction of protein-protein interactions using inference from known structures which is discussed in chapter 3.

A sequence-based prediction of protein interactions method was then used. the method uses NetPro that stores sequences of the responsible protein domains. The sequences from those known protein interactions in other species or similar proteins can serve as templates to predict an interaction of our proteins in question. Using this orthologous information and BLAST we searched for homologous interactions ($>80\%$ sequence identity) for a given protein pair [Winter et al., 2006a]. We only provided new interactions which were not confirmed before with NetPro or/and HPRD. These interactions are shown in Figure 4.13

The apoptosis map in Figure 4.14 provides only the literature confirmed interactions and three novel predicted interactions depicted by blue dotted edges which were not confirmed before with NetPro or HPRD and are found to be dysregulated in the gene expression analysis. For the three candidate genes (Bcl-2, XIAP and Survivin), we determined all known interactions then we queried the KEGG data bank to identify the involvement of these interaction partners in other pathways shown in Table 4.7.

4.4.2.3 Experimental evaluation

- **Patients and Tissues:** same as in section 4.1.2.1
- **Virtual sub array and Immunohistochemistry:** The same procedure as in case study III is used for more details see [Rückert et al., 2008].

4.4.3 Results and Discussions

4.4.3.1 Literature search

Altogether 103 genes of different parts of the cell death pathway were identified. To put the data into context, a map of the apoptotic pathway Figure 4.14 is constructed. For 54 of the 103 genes we found published data for pancreatic cancer. 82% of the total of 63 publications studied less than 5 genes or proteins at a time.

4.4.3.2 Computational evaluation of the apoptosis pathway

Figure 4.13 displays the complex nature of interactions within the apoptosis pathway. However, there exist also interconnections with other pathways. We investigated possible involvements

[3] www.molecularconnections.com

[4] www.scoppi.org

[5] www.hprd.org

of our candidate genes in other pathways to determine the effect of our new approach on cell physiology. We therefore determined all possible interaction partners of our three candidate genes by data bank search. We found 51 interaction partners for Bcl-2, 22 interactions for XIAP and 13 for Survivin. The KEGG data bank search demonstrated that through these interactions partners the three candidate genes were linked to the apoptosis and pancreatic cancer pathway. Furthermore, the genes were associated to pathways with other physiological functions like the MAPK signalling pathway, which was strongly linked to all three genes Table 4.7. We then used the protein-protein prediction method to identify potential interactions among the 103 apoptosis-associated genes. The structural folds and family assignment step (using GTD) resulted in 53 remaining assigned genes. Applying the interface conservation evaluation to possible interactions between the products of those 53 genes resulted in three novel interactions: ARTS - Apollon, p16INK -ERK and p16INK-JNK which are depicted by the blue dotted edges in Figure 4.14. A second screening with a lower evidence level was conducted for the differentially expressed genes. Furthermore, we considered interactions with more than 80% sequence identity to known interactions as homologous. The suggested new interactions of our differential expressed genes are displayed in Figure 4.13 and are depicted by red edges.

Effect of simultaneous gene silencing on apoptosis in pancreatic cancer cells For determining the effect of the silencing of the three candidate genes on apoptosis in pancreatic cancer cells is performed by measuring caspases 3 and 7 activities. The activation of caspase 3 and 7 was 5.84 times higher after the simultaneous silencing of the three genes than in controls.

Pathway analysis for the candidate genes Figure 4.14 displays interactions within the apoptosis pathway. However, there exist also interconnections with other pathways. We investigated possible involvements of our three candidate genes in other pathways to determine the effect of our new approach on cell physiology. We therefore determined all possible interaction partners of our three candidate genes by querying databases of known protein-protein interactions. The predicted interactions shows that the apoptotic pathway of cancer can choose alternative ways to stop cell death. for example the interaction between Apollon and ARTS can indicate that ARTS- an apoptosis inducer- may inhibit the activity of Apollon which is known to be an apoptosis inhibitor.

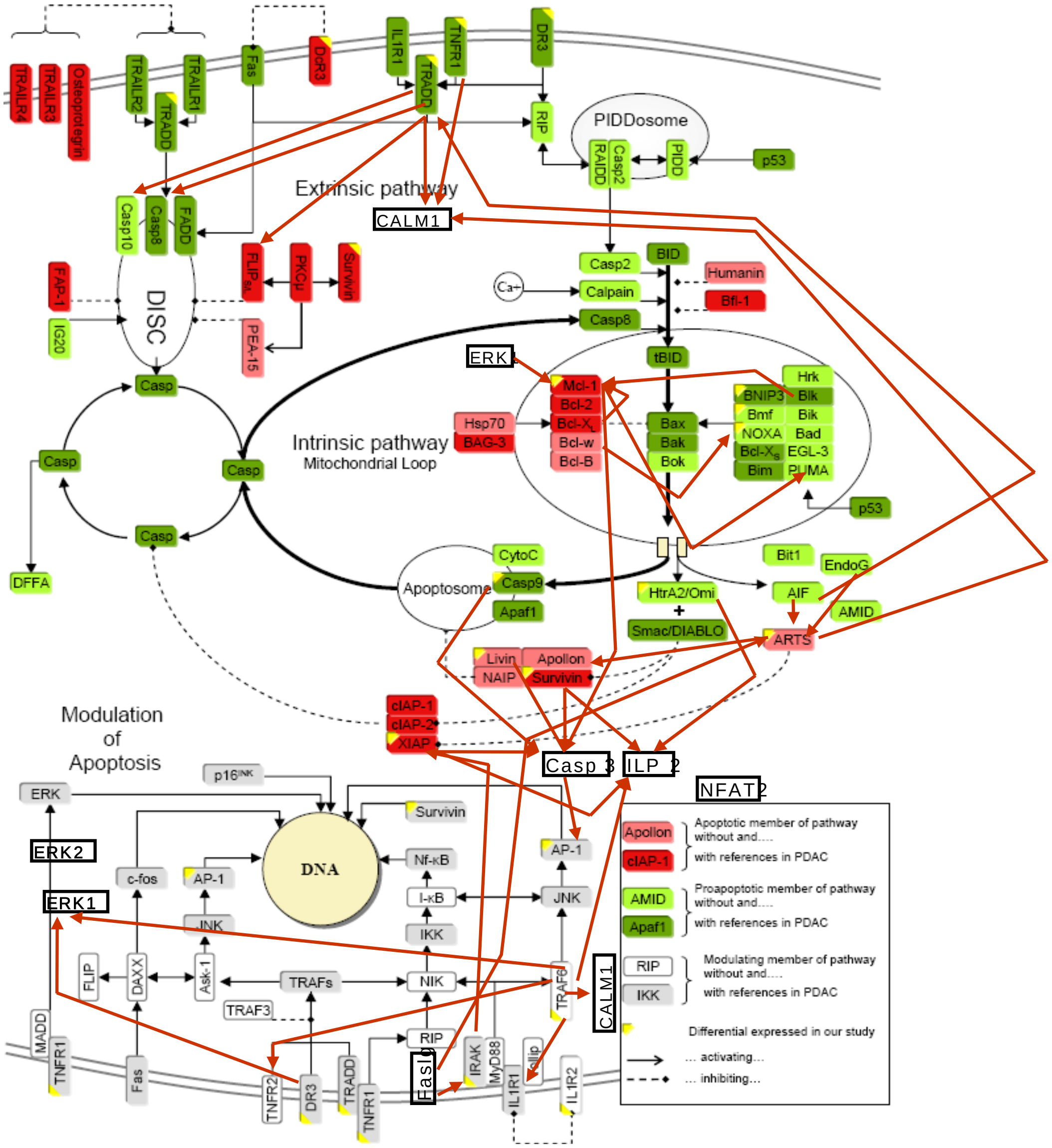


Figure 4.13: The nodes in this graph represent receptors, ligands, effectors, kinases and transcription factors while each edge describes a relation between the graph elements. The red edges indicates novel predicted interactions with sequence identity greater than 80% to known interactions from databases (HPRD, INTACT, SCOPPI) among all 103 apoptosis-associated genes. Direct apoptosis induction is shown in the upper part of the map, modulation through gene expression in the lower part. Results confirmed by our gene expression analysis are labelled by yellow corners.

.

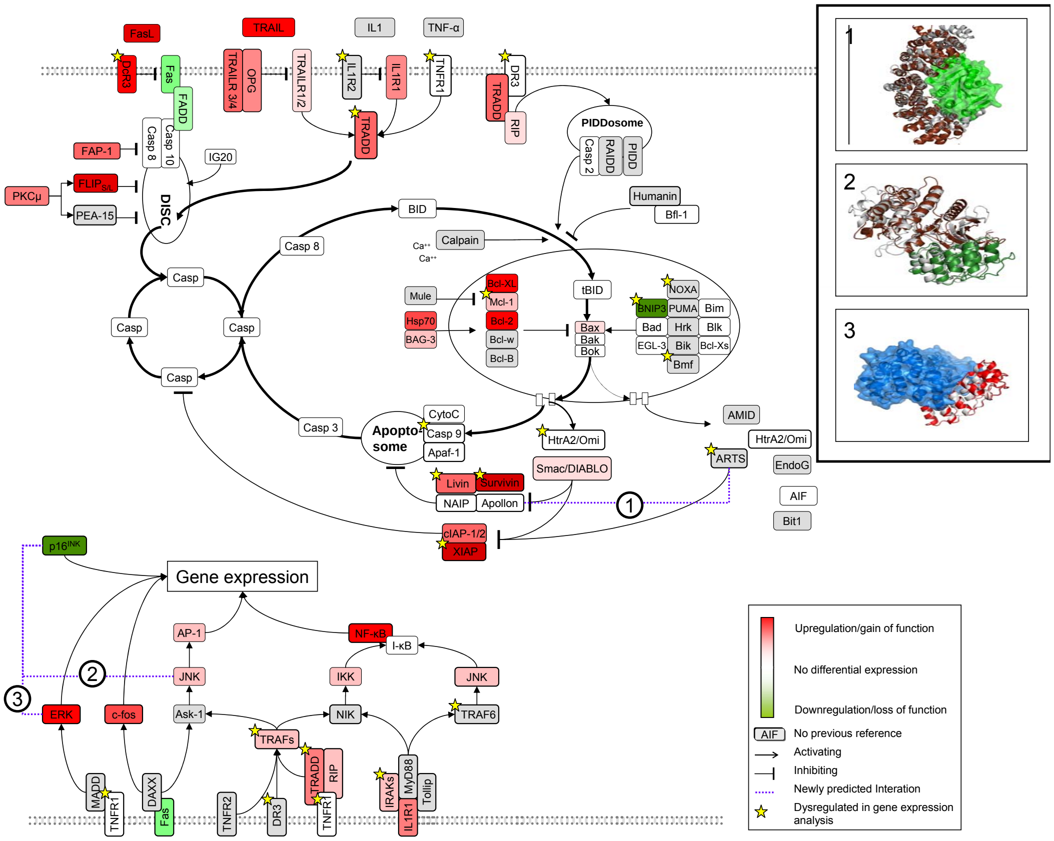


Figure 4.14: The blue edges show predicted interactions that have sequence and interaction interface identity of ≥ 30%. The structural alignment between template and interacting protein structures is below 2 Å. The structure of the three templates used in analysis for the novel interactions in Table. 4.6. The structures were used for predicting the interactions and the structural alignment of the predicted domains that belongs to the two interacting proteins. The template used for Apollon (brown) and ARTS (green) is pdb id 1qbk (grey) (1). The template used for p16INK (green) and JNK (brown) is pdb id 1blx (grey) (2). The template used for p16INK (red) and ERK (blue) is pdb id 1bi7 (grey) (3).

COUNTS	GENES	PATHWAY
13	PPP3R1, RAF1, MAPK1, TP53, PRKCA, MAPK3, MAPK8, CASP3, MAPK14, MAP3K7, TRAF6, CDC42, RASA1	MAPK signaling pathway
11	BCL2, XIAP, PPP3R1, BAX, TP53, BCL2L1, BID, CASP3, BAD, CASP7, CASP9	Apoptosis
11	BCL2, XIAP, RAF1, SRC, MAPK1, PPP1CA, PRKCA, MAPK3, MAPK8, BAD, CDC42	Focal adhesion
11	BCL2, Survivin, RAF1, BAX, MAPK1, TP53, MAPK3, MAPK8, CASP3, BAD, CASP9	Colorectal cancer
11	E2F1, RAF1, MAPK1, TP53, BCL2L1, MAPK3, MAPK8, BAD, CASP9, CDK4, CDC42	Pancreatic cancer
10	PPP3R1, RAF1, SRC, MAPK1, PRKCA, MAPK3, MAPK14, BAD, CASP9, CDC42	VEGF signaling pathway
9	BAX, TP53, BBC3, CDC2, BID, CDK2, CASP3, CASP9, CDK4	p53 signaling pathway
9	BCL2, E2F1, RAF1, MAPK1, TP53, MAPK3, CDK2, BAD, CASP9	Prostate cancer
9	BCL2, XIAP, E2F1, TP53, BCL2L1, CDK2, CASP9, TRAF6, CDK4	Small cell lung cancer
9	E2F1, RAF1, MAPK1, TP53, PRKCA, MAPK3, BAD, CASP9, CDK4	Non-small cell lung cancer
8	IRS1, RAF1, IRS2, MAPK1, PPP1CA, MAPK3, MAPK8, BAD	Insulin signaling pathway
8	RAF1, SRC, MAPK1, PRKCA, MAPK3, MAPK8, MAPK14, CDC42	GnRH signaling pathway
8	E2F1, RAF1, MAPK1, TP53, BCL2L1, MAPK3, BAD, CDK4	Chronic myeloid leukemia
7	BCL2, BAX, PSEN1, BCL2L1, CASP3, BAD, CASP7	Neurodegenerative Diseases
7	RAF1, SRC, MAPK1, PRKCA, MAPK3, MAPK8, BAD	ErbB signaling pathway
7	PPP3R1, TP53, PPP2CA, PSEN1, PRKCA, MAPK8, MAP3K7	Wnt signaling pathway
7	PPP3R1, RAF1, MAPK1, PRKCA, MAPK3, BID, CASP3	Natural killer cell mediated cytotoxicity
7	E2F1, RAF1, MAPK1, TP53, PRKCA, MAPK3, CDK4	Glioma
7	E2F1, RAF1, MAPK1, TP53, MAPK3, BAD, CDK4	Melanoma
6	RAF1, SRC, MAPK1, CDC2, PRKCA, MAPK3	Gap junction
6	MAPK1, MAPK3, MAPK8, MAPK14, MAP3K7, TRAF6	Toll-like receptor signaling pathway
6	RAF1, MAPK1, PRKCA, MAPK3, MAPK8, MAPK14	Fc epsilon RI signaling pathway
6	PPP3R1, RAF1, MAPK1, PPP1CA, PRKCA, MAPK3	Long-term potentiation
6	RAF1, MAPK1, TP53, MAPK3, BAD, CASP9	Endometrial cancer
6	E2F1, RAF1, MAPK1, TP53, MAPK3, CDK4	Bladder cancer
5	E2F1, TP53, CDC2, CDK2, CDK4	Cell cycle
5	PPP3R1, MAPK1, MAPK3, CDC42, RASA1	Axon guidance
5	SRC, MAPK1, MAPK3, MAP3K7, CDC42	Adherens junction
5	SRC, PPP2CA, PRKCA, CDK4, CDC42	Tight junction
5	RAF1, MAPK1, PPP2CA, PRKCA, MAPK3	Long-term depression
5	RAF1, MAPK1, PPP1CA, MAPK3, CDC42	Regulation of actin cytoskeleton
5	IRS1, IRS2, MAPK1, MAPK3, MAPK8	Type II diabetes mellitus
5	BCL2, BAX, TP53, BCL2L1, BAD	Amyotrophic lateral sclerosis (ALS)
5	SRC, MAPK8, CASP3, MAPK14, CDC42	Epithelial cell signaling in Helicobacter pylori infection
4	PPP3R1, SLC25A4, PRKCA, ATP2A2	Calcium signaling pathway
4	RAF1, MAPK1, MAPK3, NOTCH1	Dorso-ventral axis formation

4	RAF1, MAPK1, PRKCA, MAPK3	Melanogenesis
4	BAX, TP53, CASP3, RASA1	Huntington's disease
4	RAF1, MAPK1, MAPK3, CDC42	Renal cell carcinoma
4	RAF1, MAPK1, MAPK3, BAD	Acute myeloid leukemia
3	XIAP, UBE2D1, TRAF6	Ubiquitin mediated proteolysis
3	MAPK1, PPP2CA, MAPK3	TGF-beta signaling pathway
3	PPP3R1, CDK4, CDC42	T cell receptor signaling pathway
3	PRKCA, MAPK14, CDC42	Leukocyte transendothelial migration
3	IRS1, IRS2, MAPK8	Adipocytokine signaling pathway
3	PSEN1, CASP3, CASP7	Alzheimer's disease
3	MAPK1, TP53, MAPK3	Thyroid cancer
2	MAPK1, MAPK3	mTOR signaling pathway
2	PSEN1, NOTCH1	Notch signaling pathway
2	CASP3, CASP7	Dentatorubropallidoluysian atrophy (DRPLA)
2	PRKCA, CDC42	Pathogenic Escherichia coli infection - EHEC
2	PRKCA, CDC42	Pathogenic Escherichia coli infection - EPEC
1	HCCS	Porphyrin and chlorophyll metabolism
1	PRKCA	Phosphatidylinositol signaling system
1	BNIP1	SNARE interactions in vesicular transport
1	BCL2L1	Jak-STAT signaling pathway
1	PPP3R1	B cell receptor signaling pathway
1	BCL2	Prion disease
1	PRKCA	Cholera - Infection
1	TP53	Basal cell carcinoma

Table 4.7: Pathway analysis of the apoptosis genes and their interacting partners

4.4.4 Discussion

Previous studies in pancreatic cancer described multiple defects of apoptosis signaling at different levels of the pathway which were relevant for treatment resistance in this malignancy [Hamacher et al., 2008, Trauzold et al., 2003]. This lead to the hypothesis, that there might be a benefit in hitting multiple target genes rather than individual genes. To identify and localise appropriate candidate genes for such an approach we first constructed an apoptosis pathway map using literature search and protein-protein interaction prediction. Although the apoptosis pathway today is the best investigated pathway, the exact number of apoptosis associated genes and the exact number of interactions is still unknown. However, to our knowledge, we can for the first time present a comprehensive map of the apoptosis pathway where all interactions were validated manually and computationally with a high evidence level. Because research in apoptosis is still in progress, we have to admit that there might be more functions for the visualised proteins than depicted and that new members will possibly be discovered which we can not consider in the actual map. This hypothesis was encouraged by our protein interaction prediction, which showed three previously unknown interactions. The new interactions are an indirect sign for the complexity and the extent of this important cellular pathway. In a second step we identified and localised different clusters of defects in the cell death pathway in PDAC by literature search. It is widely agreed that pancreatic cancer cells display a conserved functionality of intracellular signal transduction mediated by executive proteins [Hamacher et al., 2008]. Upon stimulation of cell death receptors or the mitochondrial pathway an activation of caspases was detected, although not sufficient for triggering apoptosis [Vogler et al., Glazyrin et al., 2001]. This signal transduction is tightly controlled by different membrane-bound and intracellular proteins. Eukaryotic cells have developed those physiological control points to prohibit the detrimental effects on cell survival in case of inappropriate activation of programmed cell death [Westphal and Kalthoff, 2003, Fulda and Debatin, 2006]. Pancreatic cancer cells seem to abuse several of these cells own control mechanisms to stop apoptosis signaling [Vogler et al., Glazyrin et al., 2001, Satoh et al., 2001].

Downstream of the mitochondria the upregulated ”inhibitor of apoptosis proteins” (IAPs) are potent inhibitors of apoptosis by blocking both the apoptosome and the caspases [Degterev et al., 2003]. Of the eight known members of the IAP family in humans, three are upregulated in PDAC, namely Survivin, XIAP and cIAP-2 [Vogler et al., Lopes et al., 2007].

Numerous publications reported an upregulation of Bcl-2, XIAP and Survivin in pancreatic carcinoma cell lines. The pathway map makes the choice of the selection of these genes apparent since the three genes act synergistic and sequential in the flow of signal transduction. Simultaneous gene silencing showed a significant increase of apoptotic cells, of caspase activation and a significant reduction in live cells.

4.5 Limitations

The reliability of structural based interaction predictions using domains information depends on the pair of domain families involved. According to a similar study by [Aloy and Russell, 2002] that is based on the accuracy of predicted protein interactions networks using structural information, an average of 70% of interface residues are conserved in homologues complexes (cytokine/receptor 92%, signalling 89%, peptidase/inhibitor 59%, other 66%). In general estimating sensitivity and specificity for the validation of protein interaction is very difficult because there is still no comprehensive gold standard of positive (known interactions) and negative (proteins known not to interact) interaction datasets. A study by [Deane et al., 2002] using paralogs verification method (PVM) identified 40% true interactions at a 1% error rate.

For the interactions predicted from known complex structures (Table 4.1), the accuracy of structure predictions by means of Threading is crucial. Despite the fact that we filter out medium and low confidence predictions (according to confidence scores provided by the Threading method), the actual structure might still differ from the predicted one. For this reason, we compare the putative interface residues of both predicted interaction partners with the interface residues of the known complex structure used as template. We argue that a high sequence identity in the interface region favours a similar interface structure. We do not claim that these interactions are necessarily true, but we are rather confident that they provide reasonable candidates for experimental testing.

4.6 Protein structures.

The following structures were used from the Protein Data Bank: *Complex of trypsin interacting with amyloid beta-protein precursor inhibitor domain* (PDB ID 1brc) as template for modelling the TMPRSS4–TFPI2 interaction. *Crystal Structure of the Catalytic Domain of Human Complement C1S Protease* (PDB ID 1elv) to model the structure of TMPRSS4. Bovine Pancreatic Trypsin Inhibitor (PDB ID 1bpi) was used to model the structure of TFPI2. The (PDB ID 1fq1) as template modelling for the kinase inhibitor example. *Crystal structure of the Phosphorylase kinase peptide substrate complex* (PDB ID 2phk) to model the structure of Cyclin-dependent kinase 2 (CDK2). *Crystal structure of of associated Phosphatase (Kap) with a substitution of the catalytic site Cysteine to a Serine* (PDB ID 1fpz) to model the structure of Cyclin-dependent kinase inhibitor CDKN3 The BVDU structure was taken from *Crystal Structure of Thymidine Kinase from Herpes Simplex Virus Type I* (PDB ID 1ki8). For the Kinesin family member 20 A (KIF20A) *Crystal structures of mutants reveal a signalling pathway for activation of the kinesin motor ATPase* (PDB ID 1f9v) and Nicotinamide N-Methytransferase(NNMT) *Crystal structure of human pnmt complexed with skf 29661 and adohcy(SAH)*(PDB ID 1hnn).

	Gene	Description	PDAC	Gemcitabine/BVDU
1	AK3L2	adenylate kinase 3	up	down
2	AQP3	Aquaporin 3 is a water channel protein	down	up
3	AZGP1	zinc-alpha2-glycoprotein precursor	down	up
4	CASP1	Interleukin 1-beta converting enzyme isoformdelta	up	down
5	CHN1	Chimerin (chimaerin)	up	down
6	DUSP5	protein tyrosine phosphatase)	up	down
7	EPB41L4B		down	up
8	FGFR1	fibroblast growth factor receptor	down	up
9	FHL1	four and a half LIM domains 1r	down	up
10	FOXF1	forkhead box F1	up	down
11	HIST1H2AD	Histone 1, H2ad	up	down
12	IL6	interleukin 6 receptor	down	up
13	IRS1	insulin receptor substrate 1	up	down
14	KIF20A	RAB6 interacting, kinesin-like (rabkinesin6)	up	down
15	KLF4	Kruppel-like factor 4 (gut), may act as a transcriptional activator	up	down
16	LHFP	lipoma HMGIC fusion partner	down	up
17	MAOB	monoamine oxidase B	down	up
18	MT1L	metallothionein 1L	down	up
19	NNMT	nicotinamide N-methyltransferase	up	down
20	TNS3	tensin-like SH2 domain containing 1	up	down
21	PCSK6	paired basic amino acid cleaving system 4	up	down
22	PDK4	pyruvate dehydrogenase kinase, isoenzyme 4	up	down
23	PSG3	Pregnancy specific beta-1-glycoprotein 3	up	down
24	PTAFR	platelet-activating factor receptor	up	down
25	RBP1	retinol-binding protein 1, cellular	down	up
26	SCD	stearoyl-CoA desaturase	up	down
27	SLC1A4	Solute carrier family 1 (glutamate/neutral amino acid transporter), member 4	down	up
28	SMPD1	acid sphingomyelinase	down	up
29	SOD2	Superoxide dismutase 2, mitochondrial	down	up
30	SRPX	sushi-repeat-containing protein, X chromosome	down	up
31	TPM4	Tropomyosin 4	up	down
32	TMPRSS4	Transmembrane serine protease 4	down	up

Table 4.2: The overlap between the genes of the two datasets before and after treatment, where only the genes that changed expression after the combined treatment Gemcitabine/BVDU are considered.

	Protein 1	Description	up/ down	Interface conserved	Complex PDBID	Protein 2	Description	up/ down	Interface conserved
1	TFPI2	Tissue factor pathway inhibitor 2	up	62%	1brc	KLKB1	plasma kallikrein B1 precursor	up	66%
2	CCNL1	cyclin L ania-6a	down	50%	1h1s	STK17B	serinethreonine kinase 17b(apoptosis-inducing)	up	50%
3	ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	down	71%	1cc0	RHOB	Rho-related GTP-binding protein RhoB precursor (H6)	down	100%
4	SERPINB4	serine (or cysteine) proteinase inhibitor, cladeB (ovalbumin), member 3	up	52%	1k9o	ST14	suppression of tumorigenicity 14 (coloncarcinoma, matriptase, epithin)	up	63%
5	CCNL1	cyclin L ania-6a	down	50%	1h1s	TXK	TXK tyrosine kinase	down	50%
6	TFPI2	Tissue factor pathway inhibitor 2	up	61%	1c07	ST14	suppression of tumorigenicity 14 (coloncarcinoma, matriptase, epithin)	up	63%
7	CCNL1	cyclin L ania-6a	down	50%	1h1s	KIAA0536	KIAA0536 protein	down	50%
8	MICB	MHC class I polypeptide-related sequence B	up	50%	1mck	LFA-3	LFA-3(delta D2)	up	50%
9	SERPINB4	serine (or cysteine) proteinase inhibitor, cladeB (ovalbumin), member 3	up	52%	1k9o	KLKB1	plasma kallikrein B1 precursor	up	53%
10	CCNL1	cyclin L ania-6a	down	50%	1h1s	CDK6	cyclin-dependent kinase 6	up	50%
11	KLKB1	plasma kallikrein B1 precursor	up	50%	1ezx	SERPINA7	serine (or cysteine) proteinase inhibitor, cladeA (alpha-1 antiproteinase, antitrypsin), member 7	down	57%
12	CCNL1	cyclin L ania-6a	down	50%	1h1s	SNK	serum-inducible kinase	down	50%
13	ST14	suppression of tumorigenicity 14 (coloncarcinoma, matriptase, epithin)	up	71%	1ezx	SERPINA7	serine (or cysteine) proteinase inhibitor, cladeA (alpha-1 antiproteinase, antitrypsin), member 7	down	50%
14	CCNL1	cyclin L ania-6a	down	50%	1h1s	PFTK1	PFTAIRE protein kinase 1	down	50%
15	CCNL1	cyclin L ania-6a	down	50%	1h1s	TESK1	testis-specific protein kinase 1	up	50%
16	NTRK3	Neurotrophic tyrosine kinase, receptor, type 3	up	67%	1dgt	VEGF	Vascular endothelial growth factor	down	100%
17	TFPI2	Tissue factor pathway inhibitor 2	up	61%	1brc	TMPRSS4	Transmembrane protease, serine 4	up	63%

Table 4.3: Predicted protein-protein interactions among the proteins that changed expression after treatment with BVDU alone or in combination Gemcitabine/BVDU

	Gene	Name	Function	Expression
1	FOXF1	Forkhead box F1	Probable transcription activator for a number of lung-specific genesdown	down
2	FHL1	4 and half Lim1 protein	may be involved in muscle development or hypertrophy, tumor suppressor, transcriptional regulation	up
3	IFI44L	interferon induced protein 44 like	Aggregates to form macrotublar structurs (by similarity)	up
4	KIF20A	Kinesin family member 20 A	Transport of Golgi membrane	down
5	NNMT	Nicotinamide Methyltransferase	Catalyzes the N-methylation of nicotinamide and other pyridines to form pyridinium ions. This activity is important for biotransformation of many drugs and xenobiotic compounds	
6	PCSK5	proprotein convertase subtilisin/kexin type 5	Represent an endoprotease activity within the constitutive secretory pathway, with unique restricted distribution in both neuroendocrine and non-neuroendocrine	down
7	TMPRSS4	Transmembrane serine protease 4	Seems to be capable of activating epithelial sodium channel (ENaC)	up

Table 4.4: The selected genes that were of interest due to an intensive literature search that identified them as potential therapeutic markers. The respective proteins of those genes were docked using PatchDock

	Gene1	Gene2	Gene name	Comment
1	KLK10	TFPI2	Tissue factor pathway inhibitor 2	novel
2		SERPINI2	Protease inhibitor 14	novel
3	KLK6	SERPINC1	Antithrombin-III precursor(AT III)	Known
4		SERPINF2	Pigment epithelium-derived factor(PEDF)	Known
5		SNCA	Synuclein	Known
6		SERPINA3	alpha-1 antiproteinase	Known

Table 4.5: Known and potential interaction partners of KLK10 and KLK6

Protein 1	pdb protein1	IF Similarity protein 1	GS protein1	Scoppi template protein2	protein 2	pdb protein2	IF Similarity-protein2	GS protein2
TRAF2	1czy	100	99.3	1f3v	TRADD	1f3v	100	100
TRAF2	1czy	100	99.3	1f3v	TRADD	1f3v	100	100
TRAF1	1czy	100	82.3	1f3v	TRADD	1f3v	100	100
IRAK2	1muo	66.7	32	1bi7	P16INK4	1bd8	37.5	39
IRAK2	1muo	44.4	48.1	1blx	IKBA	1ikn	40.9	38.4
IRAK2	1muo	81.5	54.4	1blx	P16INK4	1bd8	40.9	38.4
IRAK3	2phk	66.7	32	1bi7	P16INK4	1bd8	62.5	42.4
IRAK3	2phk	44.4	48.1	1blx	IKBA	1ikn	45.5	42.3
IRAK3	2phk	81.5	54.4	1blx	P16INK4	1bd8	45.5	42.3
APOLLON	1bk5	41	39.3	1qbk	ARTS	1pui	42	43.9
CAPNS1	1alv	100	99.7	1kfu	CAPN2	1df0	100	100
NIK	1f3m	44.4	36.2	1blx	P16INK4	1bd8	45.5	41
NIK	1f3m	44.4	48.1	1blx	IKBA	1ikn	45.5	41
NIK	1f3m	81.5	54.4	1blx	P16INK4	1bd8	45.5	41
IKK1	1muo	66.7	32	1bi7	P16INK4	1bd8	37.5	46.8
IKK1	1muo	44.4	48.1	1blx	IKBA	1ikn	40.9	42.6
IKK1	1muo	81.5	54.4	1blx	P16INK4	1bd8	40.9	42.6
CHUK	1muo	66.7	32	1bi7	P16INK4	1bd8	75	47
CHUK	1muo	34.6	47.1	1g3n	IKBA	1ikn	30.4	49.8
CHUK	1muo	77.4	52.6	1bi8	P16INK4	1bd8	33.3	51.5
JNK	2phk	66.7	32	1bi7	P16INK4	1bd8	37.5	56.9
JNK	2phk	44.4	48.1	1blx	IKBA	1ikn	40.9	52.5
JNK	2phk	81.5	54.4	1blx	P16INK4	1bd8	40.9	52.5
JNKK	1f3m	66.7	32	1bi7	P16INK4	1bd8	37.5	48.3
JNKK	1f3m	44.4	48.1	1blx	IKBA	1ikn	40.9	46.9
JNKK	1f3m	81.5	54.4	1blx	P16INK4	1bd8	40.9	46.9
ERK	1a06	66.7	32	1bi7	P16INK4	1bd8	50	56.5
ERK	1a06	36	47.1	1g3n	IKBA	1ikn	33.3	54.9
ERK	1a06	77.4	52.6	1bi8	P16INK4	1bd8	41.7	57.6
ERK2	1f3m	66.7	32	1bi7	P16INK4	1bd8	50	58.7
ERK2	1f3m	44.4	48.1	1blx	IKBA	1ikn	36.4	54.8
ERK2	1f3m	81.5	54.4	1blx	P16INK4	1bd8	36.4	54.8
P53	1tup	50	45.4	1ycs	IKBA	1ikn	100	100
P53	1tup	33.3	42.3	1ycs	P16INK4A	1bd8	100	100
CASP8	1pau	50	30.3	1i3o	CIAP1	1g3f	77.8	66.7
CASP8	1pau	50	30.3	1i3o	CIAP2	1g3f	55.6	63.4
CASP8	1pau	50	30.3	1i3o	XIAP	1g3f	100	97.8
CASP8	1pau	50	30.3	1i3o	Survivin	1f3h	44.4	40.9
CASP8	1pau	50	30.3	1i3o	LIVIN	1c9q	66.7	65.6
CASP9	1apa	75.9	64.8	1nw9	CIAP1	1g3f	100	100
CASP9	1apa	72.4	63.7	1nw9	CIAP2	1g3f	100	100
CASP9	1apa	100	100	1nw9	XIAP	1g3f	100	100
CASP9	1apa	44.8	50.5	1nw9	Survivin	1f3h	100	100
CASP9	1apa	41.4	49.5	1nw9	Survivin	1f3h	100	100
CASP9	1apa	75.9	62.6	1nw9	LIVIN	1c9q	100	100
CASP10	1pau	75.9	64.8	1nw9	CIAP1	1g3f	31	50
CASP10	1pau	72.4	63.7	1nw9	CIAP2	1g3f	31	50
CASP10	1pau	100	100	1nw9	XIAP	1g3f	31	50
CASP10	1pau	44.8	50.5	1nw9	Survivin	1f3h	31	50
CASP10	1pau	41.4	49.5	1nw9	Survivin	1f3h	31	50
CASP10	1pau	75.9	62.6	1nw9	LIVIN	1c9q	31	50

Table 4.6: Predicted interactions among the apoptosis data set genes, where interface conservation is listed in the IF similarity protein 1/2 column and the global sequence similarity in the GS protein 1/2 column

Chapter 5

Modelling and reasoning over molecular networks: BioRevise

Designing models that demonstrate particular biological behaviours is a research field with a growing importance. These computational models aim to define the underlying theories that explain the complex manner of a biological system. A typical output of such models is an explanation of observations or abnormal behaviour of these systems [Deville et al., 2003].

Molecular networks such as metabolic pathways are complex networks that provide energy for the processes of life and synthesise new cellular material. Disorders due to defect, depletion or increase of some building blocks of these pathways can occur, leading to human diseases where the accumulated substrate may be toxic to humans or the deficiency of the product may handicap our ability to survive and function.

There is a need for identification of metabolic pathways associated with cancers that may have advantages for cancer control. Individuals who are at high risk because of pancreatic metabolism, may be candidates for dietary modification or prophylactic chemotherapy. The main idea is to remove key building blocks that cancer cells need to function instead of killing cells as with typical chemotherapy. For example tumours are known to have a common metabolic profile -high rate of Glucose uptake and macromolecule synthesis- that may confer a common selective advantage [Guffanti, 2002].

Modelling the behaviour of these networks is a challenging yet a very important task. In this research a high level representation of inhibition of enzymatic genes in metabolic pathways is presented. The model helps to identify reactions that are affected by metabolic disorders which are either genetic or acquired as a result of diet, toxins, or infections.

Our research hypothesis is that gene expression data from cancer tissues can be interpreted in terms of the metabolic pathways in which some of the co-regulated genes are involved in.

To exemplify the idea, consider that we have a network of all metabolic reactions and that this network is provided with Glucose. From the Glycolysis pathway where *Glucose* is converted to *Pyruvate*, we can infer that enzyme *Hexokinase* can convert *Glucose* into *Glucose-6-phosphate*, enzyme *Phosphoglucose isomerase* can convert *Glucose-6-phosphate* to *Fructose-6-phosphate* and *Aldolase* converts *Fructose-1,6-bisphosphate* to *Dihydroxyacetone phosphate* and *Glyceraldehyde-3-phosphate* and so on till *Pyruvate* is produced. A simplified Glycolysis

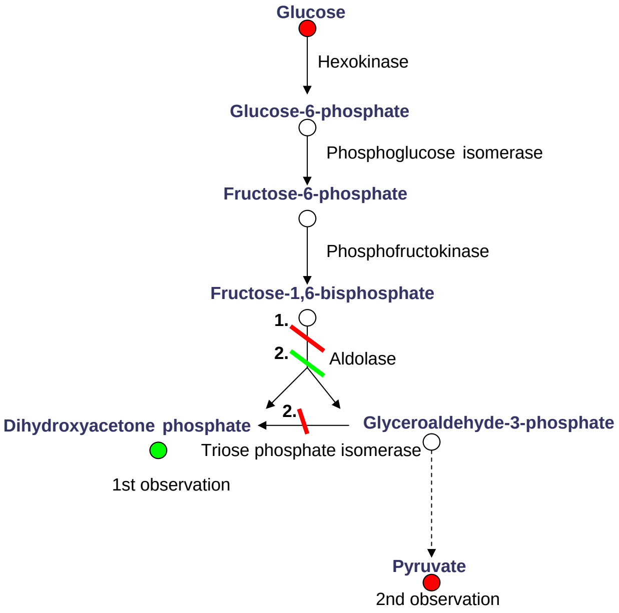


Figure 5.1: A simplified Glucose metabolic pathway where the production of Pyruvate was observed to be occurring although an intermediate metabolite (labelled with green) was not produced anymore. The picture illustrates the use of revising previous assumption to accommodate new observations while keeping the knowledge about the model consistent. The figure shows an example where a low concentration of metabolite *Dihydroxyacetone phosphate*, one explanation would be that the reaction producing the metabolite is inhibited which is indicated with a red line (solution number 1). A second observation that *Pyruvate* is still produced entails that solution 1 can not explain the two observations any more. When considerings the whole network we could show that (solution number 2) explains the low concentration of *Dihydroxyacetone phosphate* and is consistent withe other observation

pathway is sketched in Figure 5.1.

Ultimately, the network infers that the metabolite *Dihydroxyacetone phosphate* is produced. However, there is evidence that *Dihydroxyacetone phosphate* is not produced which implies that some enzymes on the path of producing this metabolite could have malfunctioned. One explanation would be that enzyme *Aldolase* is inhibited (depicted by the number 1. in Figure. 5.1), this explains that *Dihydroxyacetone phosphate* is not produced, but since *Glyceraldehyde-3-phosphate* which is also produced through the same reaction is catalysed by *Aldolase* and it is also necessary for the production of *Pyruvate* which is still present, enzyme Aldolase cannot be inhibited.

A systematic exploitation of the whole networks finally shows that inhibition of *Triose phosphate isomerase* and not *Dihydroxyacetone phosphate* (depicted by the number 2. in Figure. 5.1) explains the missing Dihydroxyacetone phosphate and is consistent with all other observed metabolites.

BioRevise a belief revision system implements the above reasoning and applies it systematically to the whole of KEGG reactions. To overcome the problem of complicated and condensed representation of the metabolic pathway maps, BioRevise provides a user friendly and interactive visualisation. BioRevise employs the Model-View- Controller, a commonly used and powerful architecture for GUIs. It makes it easier to modify either the visual appearance of the application or the underlying model rules without affecting the other.

In this chapter, BioRevise, where the inhibition of enzyme-catalysed reactions is modelled using extended logic programing, is presented. The system provides possible explanations to justify the abnormal levels of observed metabolite concentrations or dysregulation of cancer genes. These explanations are lists of the enzymes that are affected and therefore might cause certain inhibition of reactions in the metabolic pathways. The KEGG pathway database is used as the knowledge base that contains the present state of knowledge of metabolic pathways.

5.1 Introduction

Chemical reactions that take place in the cell tend to equilibrium, they are accelerated or catalysed by specialised enzymes. Enzymes are proteins that catalyse most reactions taking place in living organisms. They are considered as the main activators of different parts of the metabolic networks. Enzyme-catalysed reactions are usually connected in series, so that the product of one reaction becomes the substrate for the next. These connections of linear reactions constitutes pathways that are in turn linked to one another, forming a maze of interconnected reactions. These interconnected reactions enable the cell to survive, grow and reproduce, constituting metabolism [Alberts et al., 1998]. The inhibition of crucial enzymes can imply the interruption of a synthesis pathway and thus, to the lack of products or it can lead to the accumulation of an intermediate that is toxic when present in high concentration. If an enzyme gets inhibited, affected metabolic pathways will lead to equilibrium loss and therefore the concentration of the metabolites changes.

Pancreatic cancer is frequently associated with metabolic disorders characterised by diabetes. For example, Ariapart et al. found out that the expression of the TNF-alpha gene is upregulated in patients with pancreatic cancer and that it is involved in metabolic disorders associated with pancreatic cancer. In [Bowles et al., 2008], the authors represented a unique approach to cancer treatment in that it is one of the first to identify a metabolic pathway that can be leveraged to interrupt cancer growth. In their study Bowles et al. [2008] found that exposing the pancreatic cancer cell lines to the modified arginine deiminase enzyme inhibited cancer-cell proliferation by 50 percent. Coleman et al. hypothesised that inhibition of glucose metabolism in pancreatic cancer cells would increase cell killing via oxidative stress resulting in profound disruptions in ethiol metabolism.

This research work is motivated by the necessity to identify possible inhibited reactions caused by metabolic disorders which are either genetic or acquired as a result of introduction of toxins into the system. For this purpose we utilise a belief revision system REVISE [Damásio et al., 1997] to model inhibition of reactions in metabolic pathways. In general standard logic programs such as Prolog showed unsatisfactory treatment of negation as finite failure. In order to be able to model real biological systems two types of negations should be considered - default and explicit negations.

In default negation it is assumed in the absence of sufficient evidence that e.g. an enzyme is not malfunctioning and hence carries out the corresponding reaction. Explicit negation in the other hand states that there is evidence that e.g. an enzyme is malfunctioning, this could be a result of knocking out the enzyme. REVISE is built on top of SLX (a top-down derivation procedure) for Well Founded Semantics with eXplicit negation (WFSX) that implement these two types of negations. The top-down characterisation of WFSX relies on the construction of two types of AND trees (T and TU-trees), whose nodes are either assigned the status successful or failed. T-trees compute whether a literal is true; TU-trees whether it is true or undefined. A successful (resp. failed) tree is one whose root is successful (resp. failed). If a literal *L* has a successful T-tree rooted in it then it belongs to the paraconsistent well-founded model of the program (WFMp); otherwise, i.e. if all T-trees for *L* are failed, *L* does not belong to the WFMp. Accordingly, failure does not mean falsity, but simply failure to prove verity.

The KEGG PATHWAY database is a collection of manually drawn pathway maps representing the up to date knowledge on the molecular interaction and reaction networks for metabolism and other biological processes [Kanehisa et al., 2007]. We extract the reactions topology knowledge from the XML representation of the KEGG Pathway database and use it as part of the background knowledge (background predicates) for our belief revision model.

Modelling metabolic pathways is a nontrivial task. Besides the obvious complexity arising from the amount of data to be processed, metabolism exhibits some complex mechanisms such as negative feed back where the end product(s) of a pathway are often inhibitors of the committed step enzymes thus regulating the amount of end product made by the pathways. For a selection of current efforts in modelling metabolic networks see Table. 2.5.

We aim to analyse and model metabolic pathways in general and those associated with pancreas cancer to be able to answer queries such as “Find compounds whose concentration is directly or indirectly affected by the up/down regulation of a gene” or “Show which pathways may be affected when one or more proteins are turned off or missing”.

In this approach, we present a system that uses belief revision along with a high level representation of inhibition to reason over metabolic networks. BioRevise does not depend on kinetic information which is not available for all known metabolic reactions. BioRevise can help scientists to limit their search space of related genes to a certain observation when working with complex networks, by suggesting a concise set of affected enzymes to be checked as a result of a metabolic disorder or gene mutation effect.

5.2 Material and Methods

In the following, I will introduce some of the terminology used in the rest of the chapter.

5.2.1 Belief Revision

A belief revision occurs when a new piece of information that is *inconsistent* with the present belief system is added to that system in such a way that the result is a new consistent belief system.

5.2.2 Extended Logic Programming

Extended Logic Programming extends logic programming by integrity constraints and two types of negation: *default negation* represented by“not L” and *explicit negation* represented by “¬ L”. Where default negation gives a way of expressing a kind of negation, based on a lack of knowledge about a fact (not L is not known to be true)
Example 1: To express the assumption that the system works correctly by default we use negation by default

In a pathway *P* we expect that the end metabolite *M* is produced with a certain quantity *Q*. Therefore as long as *M* is produced with the quantity *Q* we can assume by default that *P* is normal

$$produced(M, P, Q) \leftarrow quantity(M, Q), not\ abnormal(P)$$

Explicit negation, on the other hand, allows to explicitly assert the falsity of a fact (¬ L is known to be false).

5.2.3 Integrity Constraints

Integrity constraints are used to ensure the consistency of the model e.g the model should not entail that the concentration of any metabolite is at the same time down and up.

An integrity constraint has the form

$$\perp \leftarrow L_1, ..., L_m, notL_{m+1}, ..., notL_n \ (0 \leq m \leq n)$$

where each L_i is an objective literal ($0 \leq i \leq n$), and $\perp$ stands for false. Syntactically, the only difference between the program rules and the integrity constraints is the head. A rule’s head is an objective literal, whereas the constraint’s head is $\perp$, the symbol for false. Semantically the difference is that program rules open the solution space, whereas integrity constraints limit it, as indicated in Figure 5.2 Damásio et al. [1997]).

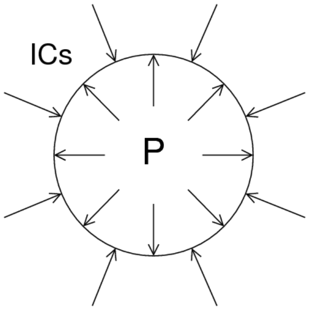


Figure 5.2: Integrity constraints closing the solution space of the program *P*.

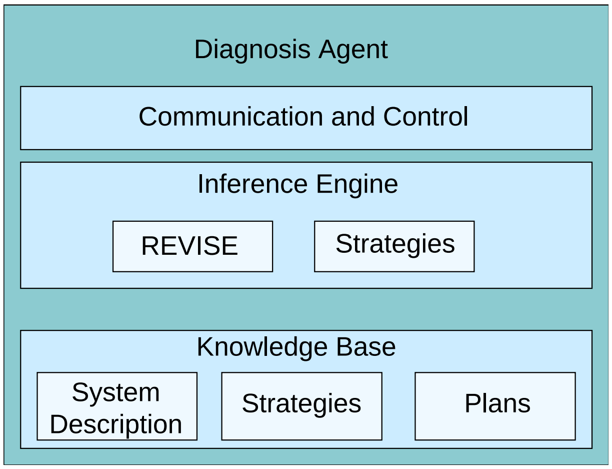


Figure 5.3: The REVISE system consists of three layers: Communication and control, a knowledge base, and an inference layer. Revise is the core of the inference machine which removes contradictions from extended Logic Programs. The strategy component employs diagnosis strategies and computes diagnosis in a process.

5.3 Resources

5.3.1 The REVISE system

REVISE is a non-monotonic reasoning system that uses belief revision to revise and remove contradictions from extended logic programs. We consider a definition of the belief revision problem that removes a contradiction from an extended logic program by modifying the truth value of a selected set of literals called revisables. The program contains as well clauses with false ($\bot$) in the head, representing integrity constraints. Any model of the program must ensure that the body of integrity constraints be false for the program to be non-contradictory. Contradiction may also arise in an extended logic program when both a literal L and its opposite $\neg$ L are obtainable in the model of the program [Lamma et al., 2001]. It is based on a top down evaluation of Well Founded Semantics with eXplicit negation (WFSX) on logic programming with explicit negation and integrity constraints. It provides two-valued revision assumptions to remove contradictions from the knowledge base. The system is embedded into an architecture for a diagnosis agent consisting of tree layers: a knowledge base, an inference layer, and on top a component for communication and control as shown in Figure 5.3. The core of the inference machine is the REVISE system, which removes contradictions from extended logic programs with integrity constraints. Additionally, there is a strategy component to employ diagnosis strategies and compute diagnosis in a process. REVISE is described in detail in [Damásio et al., 1997].

5.3.2 KEGG analysis

5.3.3 Metabolic Pathways

Pathway databases hold data on biochemical pathways and their components. Figure. 2.2 lists a number of metabolic pathways databases with specific focuses.

Metabolism has been described as modules that are divisible into relatively autonomous subunits such as the citric acid cycle or glycolysis. Metabolic pathways can be isolated from the whole cell metabolism based on specific functions. But although they might seem modular, they are interconnected, some more than the others. The highly interconnected subnetworks are called hubs, such as cell cycle, amino acid metabolism, protein synthesis, sugar metabolism, DNA metabolism, glycolysis, and tyrosine and tryptamine metabolism are good examples of hubs from KEGG [Huss and Holme, 2006].

In this work we use KEGG, a database resource for understanding higher-order functions and utilities of the biological system, such as the cell or the organism, from genomic and molecular information [Kanehisa et al., 2006]. The KEGG pathway database contains a collection of pathway maps each corresponding to a known network of functional significance [Kanehisa et al., 2004]. We extract all enzyme-catalysed reactions that are organised as connected networks. This logical representation of the network is used as the background predicates which are needed to perform the belief revision.

As of February 2009, KEGG contains 204 Homo Sapiens pathways, 111 of which are metabolic pathways containing around 2000 enzymes. Enzymes are important actors in metabolism. Most of these enzymes are very specific, they recognise and accept the mediating of only those substrates involved in one single reaction. But many other enzymes are known to be multi-functional, which means that different substrates might be bound in one or more than one active sites. In Table 5.1. an analysis of the most frequently used enzymes (enzyme families) in the KEGG metabolic pathways is presented. The table also shows the related PDAC deregulated genes and their respective KEGG pathways. KEGG provides its metabolic overviews as map illustrations and can be easier to use for the visually-oriented user.

Reactions	EC	Enzyme name	KEGG maps	Deregulated PDAC/EXP	KEGG MAPS of PDAC genes
40	1.14.13.-	Oxireductases,With NADH or NADPH as one donor, and incorporation of one atom of oxygen	17	COQ6/Up	hsa00130
31	2.4.1.-	Hexosyltransferases	9		
24	4.2.1.17	Enoyl-CoA hydratase	11	HSD17B4/Down	hsa00150
23	4.2.1.-	Hydro-lyases	11		
22	2.3.1.-	Transferring groups other than amino-acyl groups	14		

18	1.2.1.3	aldehyde dehydrogenase (NAD+)	16	ALDH1A2/Down, ALDH1A1/Down	hsa00010 hsa00053 hsa00071 hsa00120 hsa00280 hsa00310 hsa00330 hsa00340 hsa00380 hsa00410 hsa00561 hsa00620 hsa00640 hsa00650 hsa00903
18	4.1.1.-	Carboxy-Layases	15		
17	2.3.1.16	acetyl-CoA C-acyltransferase	5		
16	1.14.-.-	Acting on paired donors, with incorporation or reduction of molecular oxygen	7		
15	2.7.1.69	protein-Npi-phosphohistidine-sugar phosphotransferase	5		
14	1.13.11.-	Oxidoreductases,With incorporation of two atoms of oxygen	8		
13	2.3.1.85	fatty-acid synthase	2		
13	2.1.1.-	Methyltransferases	8	HRMT1L2/Up	hsa00150 hsa00340 hsa00350 hsa00380 hsa00440 hsa00450 hsa00626
12	1.2.1.-	Oxidoreductases,With NAD+ or NADP+ as acceptor	8		
12	1.1.1.35	3-hydroxyacyl-CoA dehydrogenase	7		
12	3.1.3.5	5'-nucleotidase	3	NT5E/Up	hsa00230 hsa00240 hsa00760
12	2.4.99.-	Transferring other glycosyl groups	2		
12	1.1.1.-	Oxireductases,With NAD+ or NADP+ as acceptor	12		
11	2.4.1.17	glucuronosyltransferase	4	UGT2B28/Down	hsa00040 hsa00150 hsa00500 hsa00860
11	3.2.1.22	a-galactosidase	4		
11	2.7.4.6	nucleoside-diphosphate kinase	2	NME1/Up	hsa00230 hsa00240

11	1.4.3.4	amine oxidase (flavin-containing)	5	MAOB/Down	hsa00260 hsa00330 hsa00340 hsa00350 hsa00360 hsa00380
11	4.2.1.80	2-oxopent-4-enoate hydratase	9		
11	1.1.-.-	Acting on the CH-OH group of donors	6		
10	1.1.1.50	3alpha-hydroxysteroid dehydrogenase	3		
9	6.2.1.-	Acid-Thiol Ligases	5		
8	1.2.1.5	aldehyde dehydrogenase [NAD(P)+]	4		
8	2.4.1.69	galactoside 2-a-L-fucosyltransferase	3		
7	5.3.3.1	steroid DELTA-isomerase	2		
6	1.2.1.18	malonate-semialdehyde dehydrogenase (acetylating)	4		

Table 5.1: An analysis of the most frequently occurring enzymes in KEGG. The larger/smaller the number of KEGG pathways those enzymes act on, the more general/specific these enzymes are. Relative PDAC enzymes with their respective expression are shown along with the KEGG pathways they are associated to.

5.3.4 BioRevise architecture: Model-View-Controller

The Model View Controller paradigm is a classic design pattern that has been pursued for a clear design which separates objects within an interactive application into one of three categories models for maintaining data, views for displaying all or a portion of the data, and controllers for handling events that affect the model or view(s).

A *model* in this paradigm is a class which originates in a specific domain. It is an abstraction of a domain specific entity and has no knowledge about the graphic user interface (GUI). The representation of the *model* as GUI element is called *view*. A *view* can be seen as a wrapper around the *model*, which is capable of displaying a subset of the data that is encapsulated in the *model*. Each view has an associated *controller*. A *controller* is responsible for all possible actions that are defined in the *view* concerning the associated *model*. A *model* can have multiple *views* [Veit and Herrmann, 2003]. The *model* does not only captures the state of the system, but also how the system works. The model view controller separates the user interface from the core application data and functionality. With this separation one of the components can change without requiring changes in the others components. This is an alternative to the traditional (input, processing, output) applications. In the model view controller perspective the keyboard and mouse inputs are handled by the *controller* that makes the proper connection with the other two components, the *view* and the *model*. The *model* is able to change state, answer about its state and manage the needed data structures. Presenting the data related to the state or changes in the *model* is the role of the *view* component in the model view controller.

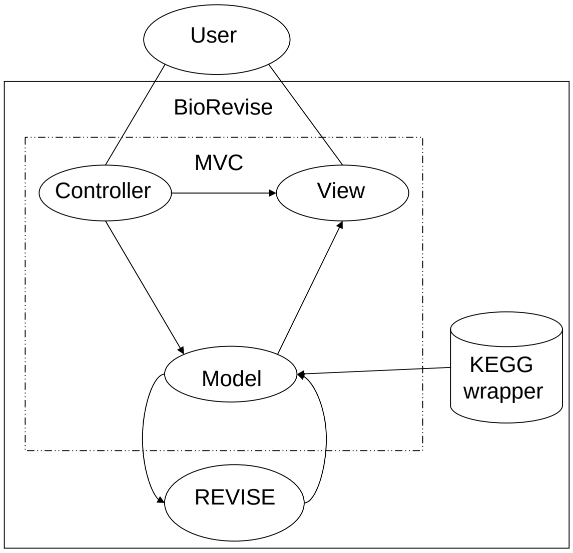


Figure 5.4: BioRevise system work flow. The connections between the main components of the BioRevise system.

5.4 BioRevise prototype work flow

The starting point of the BioRevise system is the extraction of the background predicates from the KEGG database. The background knowledge consists of the background predicates and the metabolites concentrations observed (observable predicates). It is used by the *model* to write the system input files. These observations are marked in the maps of the metabolic pathways shown by the *view*.

BioRevise takes as input the extracted background knowledge and then uses it to revise the inhibition model, and as output it produces lists of possible inhibited enzyme-catalysed reactions that explain the abnormal levels of the observed metabolites concentration or deregulated enzymes. The background knowledge, program (definitional knowledge), and integrity constraints representing the metabolism are modelled using Extended Logic Programs while the visualisation is implemented using Java. The *view* shows the corresponding enzyme-catalysed reactions in the maps using the extracted coordinates of the enzymes. Through the *controller* the user is also able to *zoom in* from the maps of the different metabolic families to the metabolic pathways maps with the inhibited reaction. Figure 5.4 shows the connection between the different components of the system.

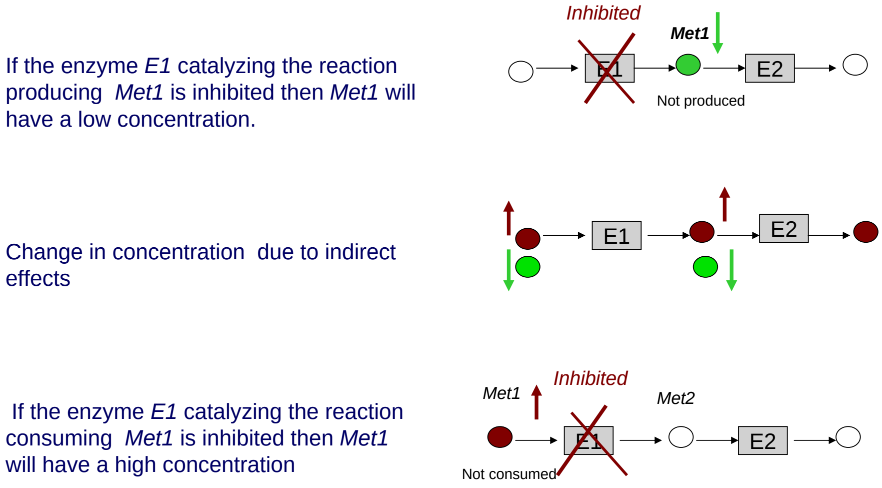

If the enzyme E1 catalyzing the reaction producing Met1 is inhibited then Met1 will have a low concentration.
Inhibited
Met1
E1
E2
Not produced
Change in concentration due to indirect effects
E1
E2
If the enzyme E1 catalyzing the reaction consuming Met1 is inhibited then Met1 will have a high concentration
Met1
Inhibited
Met2
E1
E2
Not consumed

Figure 5.5: The knowledge modelling: The first step is used to represents the fact that if the reaction that produces a product *Prod* is inhibited, this will cause a decrease in the concentration of the product *Prod*. The second step represents the changes on the concentration caused through indirect effects, where a metabolite *Prod* can have *down*/*up* concentration due to the fact that some other substrate metabolite *Sub*, that produces *Prod* has a decrease/increase of concentration respectively (even when the reaction is apparently not inhibited). The third step is used when an enzyme catalysing a reaction that consumes a substrate *Sub*, this will cause an increase of *Sub*.

5.4.1 REVISE

REVISE is the core of the BioRevise system. It is used to perform belief revision over the knowledge representation of the metabolic pathways, it is integrated in the *model* component of the model view controller. REVISE uses three input files, the first file consisting of the observables (metabolite concentration levels) provided by the user, the second file contains the network topology rules and the third file contains the knowledge modelling describing the inhibition behaviour and the integrity constraints.

5.4.2 Knowledge Modelling

The knowledge modelling is the computational representation of the metabolism. By capturing the knowledge and behaviour of the components of the metabolism system, some inferences can be made about changes in the state of the BioRevise system based on the new states due to abnormal observations obtained by the system. In this work we concentrate in modelling of inhibition. The modelling presented here is based on a previous work developed by [Tamaddoni-

Nezhad et al., 2004]. The model of the metabolism can separate two disjoint sets of predicates: the *observable* predicates and the *abducible* predicates. The model can be *incomplete* in its description. To complete the description, the new information given by the observation can be used. The basic assumption is that all the incompleteness of the model can be isolated in its abducible predicates. The *background* predicates are auxiliary relations linking observable and abducible information. For this purpose a logic program is required which models how the concentration of metabolites is related to inhibition of enzymes [Tamaddoni-Nezhad et al., 2004]. To represent the changes on the metabolites concentration, the *observable* predicate *obs*/2 is used:

$$obs(Metabolite, Concentration)$$

e.g.

$$obs('Pyruvate', up)$$

This predicate encodes the observations made by the user, where variable *Concentration* can be either *up* or *down*. The relational representation of metabolic networks that form the metabolism is represented by the background predicates *reaction*/4:

$$reaction(Sub, Enz, Prod).$$

representing the fact that the enzyme-catalysed reaction occurs in a direct path from one node to the other, where *Sub* is a set of substrates and *Prod* is the product they produce. The product of one reaction becomes the substrate of another. For example the predicate:

$$reaction(['Acetate'], '6.2.1.1', 'Acetyl-CoA').$$

represents the enzyme-catalysed reaction between *Acetate* and *Acetyl-CoA* catalysed by the enzyme *Acetate-CoA ligase* (EC:6.2.1.1) in the ”Glycolysis/Gluconeogenesis” pathway.

The incompleteness of the model resides in the lack of knowledge of which metabolic reactions are adversely affected in the event of a metabolic disorder Tamaddoni-Nezhad et al. [2004].

To predict the inhibition of one reaction and complete the model, the abducible predicate *inhibited* predicate is used:

$$inhibited(Enz, Sub, Prod)$$

encoding the fact that the enzyme-catalysed reaction producing the product *Prod* from the substrate *Sub* is inhibited because of the inhibition of the enzyme *Enz*, due to a metabolic disorder of the system. For example:

$$inhibited('6.2.1.1', 'Acetyl-CoA', 'Acetate')$$

which captures the hypothesis that the reaction from *Acetyl-CoA* to *Acetate* that is normally catalysed by the enzyme *6.2.1.1* is inhibited due to a certain defect in this enzyme.

For describing the behaviour of the system due to inhibition of a certain reaction, the following rules are used:

$$\begin{aligned} &concentration(Met, down) \leftarrow \\ &\qquad reaction(Sub, Enz, Prod), \\ &\qquad member(Met, Prod), \\ &\qquad inhibited(Enz, Sub, Prod). \end{aligned} \tag{5.1}$$

$$\begin{aligned} &concentration(Met, up) \leftarrow \\ &\qquad reaction(Sub, Enz, Prod), \\ &\qquad member(Met, Sub), \\ &\qquad inhibited(Enz, Sub, Prod). \end{aligned} \tag{5.2}$$

$$\begin{aligned} &concentration(Met, down) \leftarrow \\ &\qquad reaction(Sub, Enz, Prod), \\ &\qquad member(Met, Prod), \\ &\qquad not\ inhibited(Enz, Sub, prod), \\ &\qquad concentrationsOne(Sub, down). \end{aligned} \tag{5.3}$$

$$\begin{aligned} &concentration(Met, up) \leftarrow \\ &\qquad reaction(Sub, Enz, Prod), \\ &\qquad member(Met, Prod), \\ &\qquad not\ inhibited(Enz, Sub, prod), \\ &\qquad concentration\_all(Sub, up). \end{aligned} \tag{5.4}$$

We consider all substrates that are used as substrates, but not a product of any reaction as “input” metabolites.

$$\begin{aligned} &input(Sub) \leftarrow \\ &\qquad reaction(Sub, Enz, Pro), \\ &\qquad not\ reaction(\_, \_, Sub). \end{aligned}$$

We then assign the concentration "up" to all "input" substrates.

$$\begin{aligned} concentration(Sub, up) \leftarrow & \\ & not\ concentration(Sub, down), \\ & input(S). \end{aligned}$$

These rules describe the model in a simplified way, considering the fact that some metabolic disorders can change the concentration of metabolites by the inhibition of enzymes catalysing enzyme-catalysed reactions. An illustrative description of the rules is shown in Figure 5.5. The first rule (5.1) is used to represents the fact that if the reaction that produces a product *Prod* is inhibited, this will cause a decrease in the concentration of the product *Prod*. The second rule (5.2) is used when an enzyme catalysing a reaction that consumes a substrate *Sub*, this will cause an increase of *Sub*. The third and fourth rules (5.3 and 5.4) represents the changes on the concentration caused through indirect effects, where a metabolite *Prod* can have *down*/*up* concentration due to the fact that some other substrate metabolite *Sub*, that produces *Prod* has a decrease/increase of concentration respectively (even when the reaction is apparently not inhibited). The auxiliary rules for *member*, *concentrationOne* and *concentrationall* deals with sets operations. *concentrationOne* ensures that for a product to be down, at least one of the members of the substrates set producing this product is down and *concentrationall* ensures that for a product to be up, all members of the substrates producing this product has to be up.

The integrity constraints of the model captures several *validity requirements* that must be satisfied by the abducible information of *inhibited*/3. In our model to express the fact that the concentration of a metabolite can not be *up* and *down* at the same time the following integrity constraints are used:

$$\leftarrow concentration(Metabolite, up), concentration(Metabolite, down).$$

5.5 Implementation

The model view controller model is implemented jointly with Vasco Pedro which was part of his master thesis work under my supervision.

5.5.1 Model

The data structures populated by the KEGG extractor with the knowledge base of the BioRevise system are used by the *model* to write the REVISE observables input file. The file contains the observations marked by the user corresponding to the new concentration of metabolites. The concentration observed are either *up* or *down*. The REVISE system is then used to read the other input files and generate the output file containing the possible explanations. These explanations are the possible inhibited enzyme-catalysed reactions caused by the modification in the concentration of the metabolites. Using these results the *model* updates the inhibition state of the enzymes in the data structures from *not inhibited* to *inhibited*, in order to also display it in the *view*.

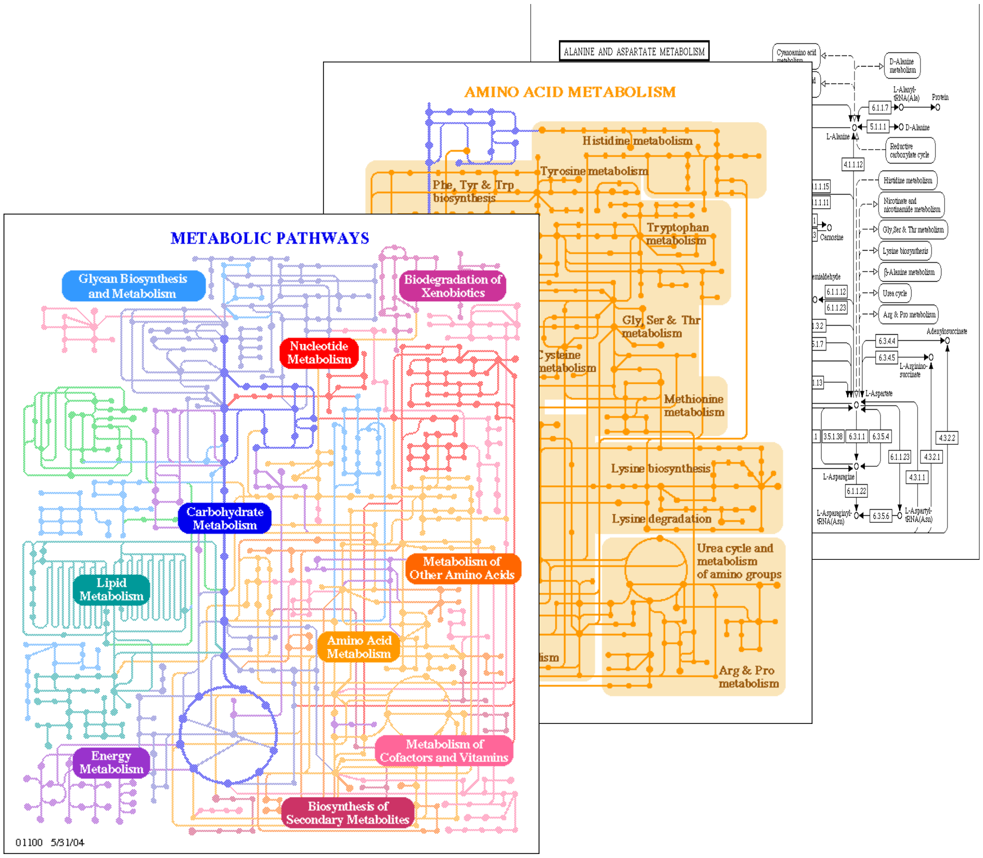
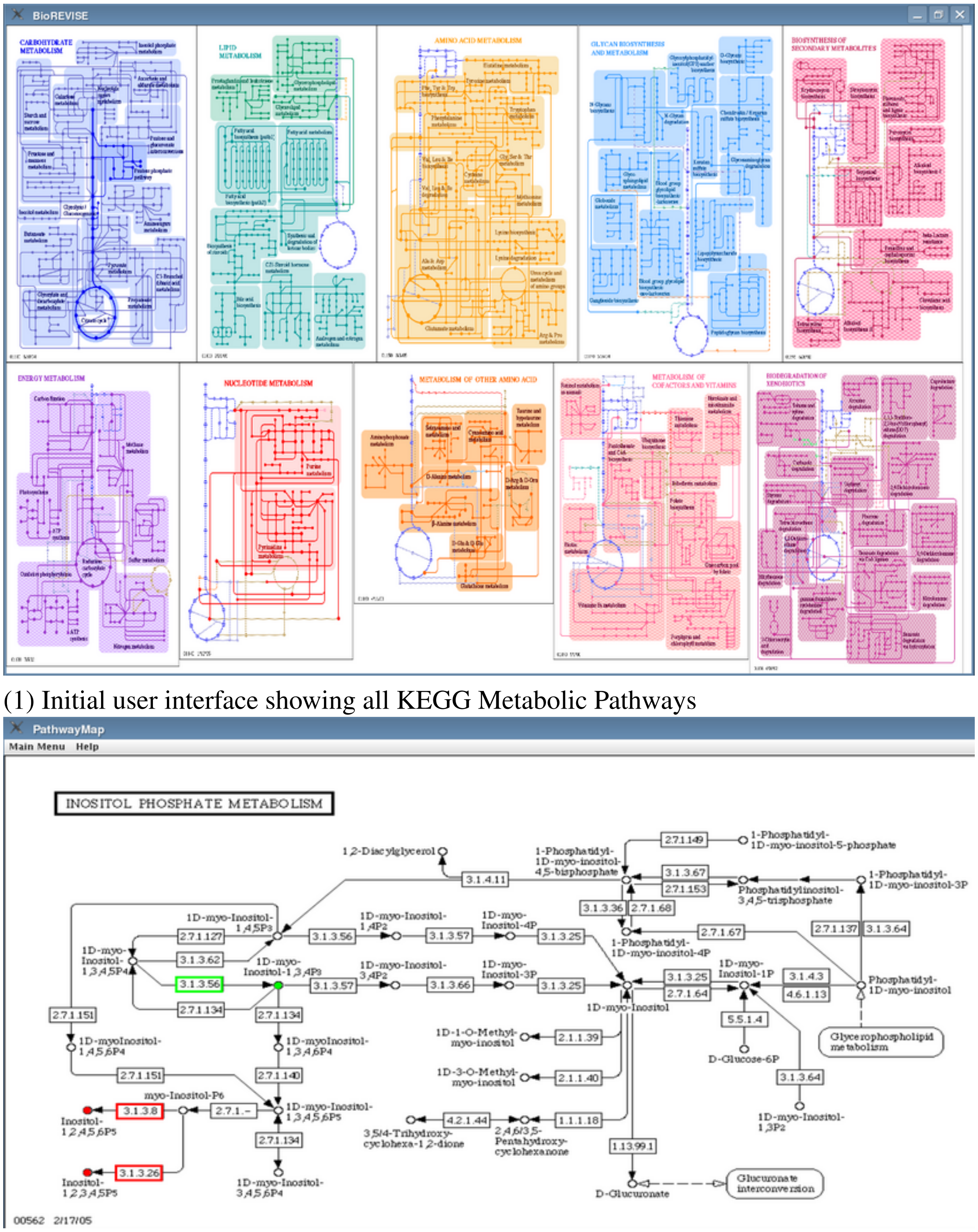


Figure 5.6: The three abstraction levels of the KEGG maps.

5.5.2 Visualisation

Due to the complex representation of metabolic networks (e.g. Boehringer map), finding enzymes that are affected as a side effect of an inhibition of certain reaction is like searching for a needle in a hey sack. The metabolic pathway maps are build in three different levels of abstraction. The first level is the map representation of all the metabolism network, the second is the collection of maps within the different groups of metabolism, and the third level contains the maps of the metabolic pathways, which are the most detailed maps also representing the pathway reactions and their components shown in Figure 5.6. To perform the visualisation of the metabolic pathway networks and to simplify finding the enzyme catalysed-reactions, the maps from KEGG are used.

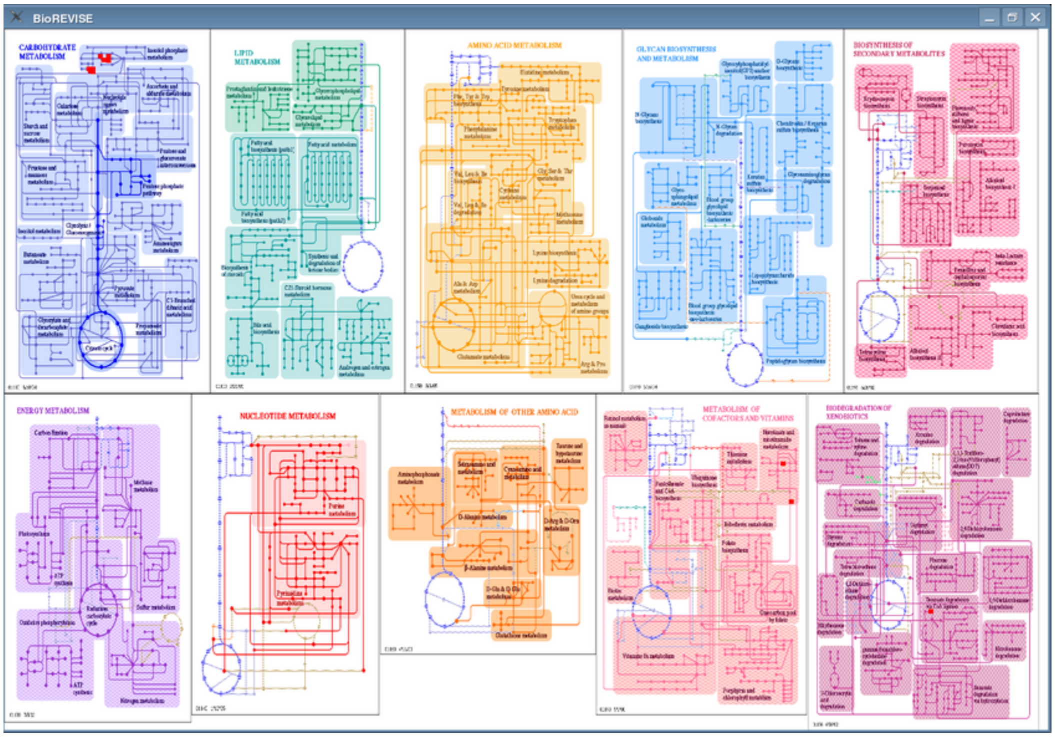
The first picture of the metabolism is composed of the maps that represent the different groups of metabolism, biosynthesis and biodegradation in KEGG. The generated picture is outlined by putting together the maps according to the coordinates from the *xml* files. The coordinates of the *xml* files were manually built to present all maps in the same picture. By clicking on any of the metabolic pathways in the first picture, it is possible to see the most detailed maps from KEGG. These maps contain the representation of the enzyme-catalysed



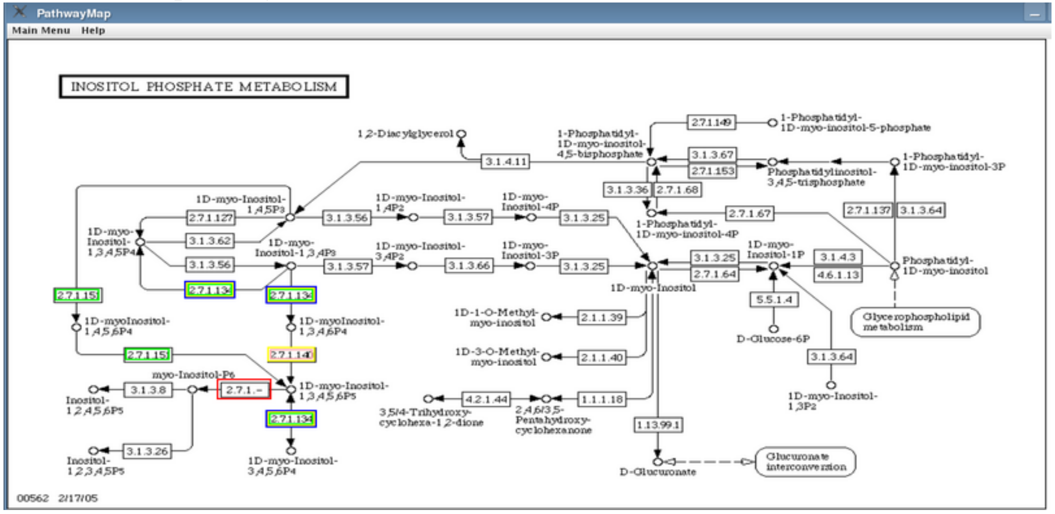
(1) Initial user interface showing all KEGG Metabolic Pathways

(2) Assigning concentration levels to Enzymes/Metabolites (red for up and green for down)

Figure 5.7: (1) First window showing maps of the different groups of metabolism, biosynthesis and biodegradation in KEGG. (2) The concentration or regulation levels of the metabolites and enzymes are marked using different colours, green for down and red for up.


BioREVISE
CARBOHYDRATE METABOLISM
LIPID METABOLISM
AMINO ACID METABOLISM
GLYCAN BIOSYNTHESIS AND METABOLISM
BIOSYNTHESIS OF SECONDARY METABOLITES
ENERGY METABOLISM
NUCLEOTIDE METABOLISM
METABOLISM OF OTHER AMINO ACID
METABOLISM OF COFACTORS AND VITAMINS
BIODEGRADATION OF XENOBIOTICS

(3) The red squares indicating the reactions affected as a result for the abnormal regulation levels of the input Enzymes/Metabolites



(4) Highlighted inhibited reactions with the enzymes that catalyse the reactions.

Figure 5.8: (3) The visualisation of the inhibited reactions, the effected reactions are marked in the first window in the region that corresponds to a metabolic pathway. (4) Inhibited reactions with the enzymes that catalyse the reaction are highlighted with a red, blue, green or pink rectangles, corresponding to the first, second, third or fourth solution respectively at the second window.

reactions that constitute the metabolism. In the first two windows see Figure 5.6 the user can perform the operations of *zooming in* and *out*, as well as *translate* the picture by moving it up or down. It is also possible to mark the observations made by the user in this second window. This corresponds to the change of the metabolites concentration to *up* and *down*. The concentration of the metabolites is marked using different colours, green for down and red for up as illustrated in Figure 5.7. For the visualisation of the inhibited reactions, the affected reactions are marked in the first window in the region that corresponds to a metabolic pathway, shown in Figure 5.8. At this stage, the user is able to see all the inhibited reactions in the metabolism. The user can then visualise the maps of the metabolic pathways in the second window to see each of the enzyme-catalysed reactions which are inhibited. To show these inhibitions on the second window, the enzymes that catalyse the reaction are highlighted with red, blue, green or pink rectangles, corresponding to the first, second, third or fourth solutions respectively.

5.5.3 Controller

The *controller* performs the connection between the *model* and the *view*, by reacting to the mouse and keyboard events. For each event the *controller* performs the corresponding data update in the data structures, used by the *model* or the *view*. Other data related to the *zoom* and *translation* operations made in the maps, is also updated.

5.6 Results and Discussion

We provide a system that uses belief revision to model reaction inhibition in metabolic pathways. Given abnormal concentration levels of metabolites, the system will reason over the KEGG network to show all reactions and pathways affected due to the metabolic disorder that caused the metabolites levels to behave abnormally. The system used a high level representation of inhibition, which makes it independent of detailed kinetic information of metabolism modelling.

We demonstrate the use of the BioRevise system by applying it to the two examples presented in section 5.6.1 and 5.6.2.

5.6.1 Metabolic disorder example: Glycogen storage disease

Enzyme	EC Number	Disease	Metabolites
Glucose-6-phosphatase.	3.1.3.9	Glycogen storage disease I	lactic acid and Glycogen levels increase, glucose decrease
Polyribonucleotide nucleotidyltransferase (PNPase)	2.7.7.8	Lymphopenia and ISCD	serum urate
Glucose-6-phosphatase(G6Pase)	3.1.3.9	Glycogen Storage Disease	uric acid level increase, glycogen level decrease
Xanthine oxidoreductase(XOR)	1.16.3.1	Xanthinuria	uric acid level decrease and xanthine level increase
Phenylketonuria	1.14.16.1	Phenylalanine hydroxylase	phenylalanine level increase.
Alkaptonuria Ochronosis	1.13.11.5	homogentisic acid oxidase phenylalanine or tyrosinelevel increase .	homogentisic acid level increase

Tyrosinemia I	3.7.1.2	enzyme fumarylacetoacetate hydrolase	amino acid tyrosine can not be broken down
Histidinemia		histidase	histidine levels increasee in blood and urocanic acid in blood, urine, and skin cells
Maple syrup urine disease	1.2.4.4	branched-chain alpha-keto acid dehydrogenase	leucine, isoleucine, and valine levels increase
Propionic acidemia	6.4.1.3	propionyl-CoA carboxylase	propionyl-CoA level increase
Methylmalonic acidemia		methylmalonyl-CoA mutase	defect in the conversion of methylmalonyl-coenzyme A (CoA) to succinyl-CoA
Isovaleric acidemia,3	1.3.99.10	isovaleric acid-CoA dehydrogenase	isovaleric acid level increase
Methylcrotonyl-CoA carboxylase deficiency		3-methylcrotonyl-CoA carboxylase.	leucine level increase

Table 5.2: A number of known enzymes deficiency disorders and the enzymes responsible for the disease.

For metabolic diseases there are 666 entries from OMIM, 43 in BRENDA, and 110 collected in Wikipedia. Table 5.2. lists a number or known enzyme deficiency disorders. The EC column contains the enzymes responsible for the disease and the metabolites column contains the observed change of metabolite concentration levels.

Glycogen storage disease type I (GSD I) is a metabolic disorder that is caused by the deficiency in the glucose-6-phosphatase enzyme. This deficiency impairs the ability of the liver to produce free glucose from Glycogen and from Gluconeogenesis. Glycogen and Gluconeogenesis are the two principal metabolic mechanisms by which the liver supplies Glucose to the rest of the body during periods of fasting. Reduced Glycogen breakdown results in increased Glycogen storage in liver and kidneys, causing enlargement of both. An obvious symptom of the GSD type I is the inability to maintain adequate blood glucose levels during fasting which results from the combined impairment of both Glycogenolysis and Gluconeogenesis. According to the BioRevise model, there are many possible hypothesise that can explain the abnormal concentration levels of the observed metabolites. However the system provides only the most comprehensive and short explanations. We used the observed values of the metabolites caused by the GSD I (the Glucose down, Pyruvate lactate up) as input for the BioRevise system.

BioRevise could identify that the inhibition of the enzyme glucose-6-phosphatase (EC:3.1.3.9) as a possible explanation for the abnormal levels of concentration of Glucose, which according to literature is known to be the main reason for this disease. This example is a proof of concept and shows the usability of the BioRevise system.

5.6.2 Reasoning over pancreatic cancer associated metabolic disorders

Rapidly growing cancer cells typically have higher glycolysis rate than those of their normal tissues of origin. One way to explain this phenomenon is by following the hypothesis of Otto Warburg (1930), which claims that cancer is primarily caused by dysfunctionality in mitochondrial metabolism, rather than because of uncontrolled growth of cell.

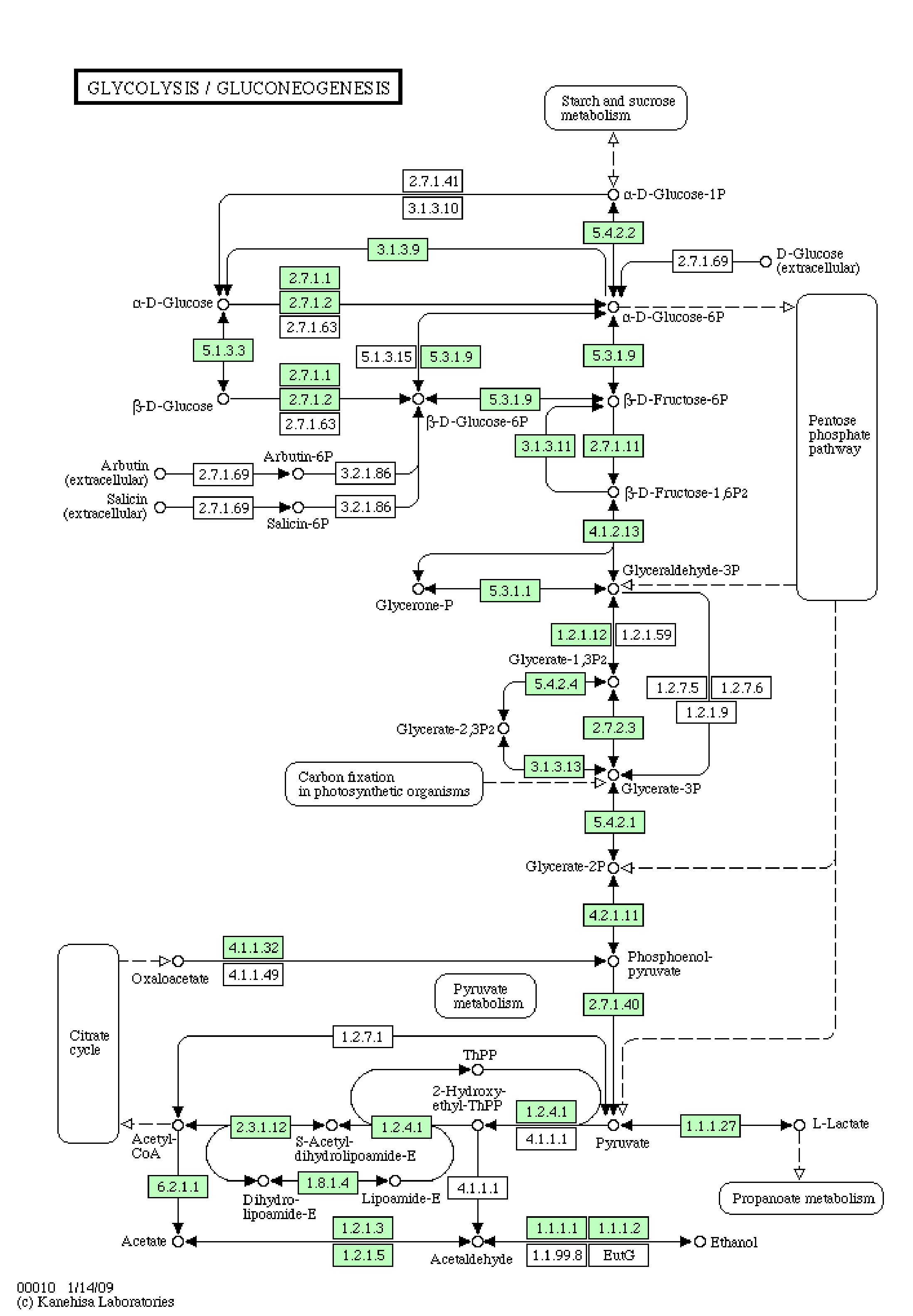


Figure 5.9: KEGG Glycolysis/Gluconeogenesis reference map, showing in green the involved human enzymes

KEGG provides two Glucose metabolism pathways: the main one is the Glycolysis / Gluconeogenesis pathway shown in Figure 5.9 and an alternative pathways Pentose phosphate pathway. Glycolysis is the process of converting glucose into Pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). We used BioRevise to generate possible explanations of the effect of inhibiting the Glucose pathway along with taking into account the change of regulation of the Glucose pathway enzymes in the PDAC data. For this step we propagate the level of regulation of PDAC enzymes associated with the Glycolysis / Gluconeogenesis pathway to the metabolites produced by the reaction catalysed by these enzymes. As an example, if in the reaction $reaction(['Acetate'],'6.2.1.1','Acetyl - CoA')$ and we know that the enzyme *6.2.1.1* is down regulated we assume that the product *Acetyl-CoA* has low concentration.

As input we provide the system with the observations that the genes coding for the enzyme Fructose-bisphosphate aldolase (ALDOB) (EC:4.1.2.13) is down regualted and for the enzyme glucose-6-phosphate isomerase (GPI) (EC:5.3.1.9) is up regulated from the PDAC data set. The BioRevise provides the following solutions as explanations

```
Solution1=[[inhibited(’2.7.1.11’,
           [’beta-D-Fructose 6-phosphate’],
           [’beta-D-Fructose 1,6-bisphosphate’])], []] ;

Solution2=[[inhibited(’4.1.2.13’,
           [’beta-D-Fructose 1,6-bisphosphate’],
           [’(2R)-2-Hydroxy-3-(phosphonooxy)-propanal’]),
            inhibited(’4.1.2.13’,
           [’beta-D-Fructose 1,6-bisphosphate’],
           [’Glycerone phosphate’])], []] ;
```

The explanation provided by solution one is an inhibition of an upstream reaction of the path to the reaction catalysed by (ALDOB). This solution explains the two observations mentioned above. The second solution shows that the reaction producing *Glycerone phosphate* and *(2R)-2-Hydroxy-3-(phosphonooxy)-propanal*, has to be inhibited also to explain the observations. This kind of information is instructive to future experiments where concentration levels of some metabolites can be tested in order to validate or refute the second solution. The use of the BioRevise GUI make is easy to points out inhibition of reactions that are not in the same pathway but uses the same enzyme or produce the same metabolite. In this way BioRevise can provide explanation that might guide us to understand the mechanism of action of complex diseases by studying the pathways that are mainly affected by cancer.

5.6.3 Limitations

The current version of BioRevise can not handle the whole KEGG metabolic pathway network. One reason is due to the possible loops that can be produced by negative and positive feed back mechanisms, where for example the end product(s) of a pathway are often inhibitors of the first reaction enzymes thus regulating the amount of end product made by the pathways. This could be handled by eliminating unimportant loops by reducing the KEGG pathways into simpler reference maps that does not contain such loops. Another limitation is that the enzyme-catalysed reactions from the metabolic pathways can be reversible, in order to compensate the variation of concentration in one of the metabolites. However the current prototype does not consider the possibility of the reversed enzyme-catalysed reactions that can occur to compensate the changes in the metabolites concentrations.

Chapter 6

Summary and discussion

6.1 Open research problems revisited

6.1.1 Open problem: Can the use of protein-protein interaction data enhance knowledge discovery from gene expression data analysis to reveal potential markers and therapeutic targets for Pancreatic cancer

.

To provide solutions to this problem, we developed a method that enables gene expression analysis within the biological context of protein-protein interactions and pathways. High throughput methods such as microarray enabled large scale analysis of gene expression data that produces thousand of genes. However, analysing of such data for reaching a functional understanding of ongoing processes is a major challenge, in particular when thinking in terms of a systems biology driven analysis. Further investigations of such data is hampered by the fact that except for the sequence rather little is known about those genes. Structure recognition is still lagging behind the out put of proteins sequence data. Protein interactions provide an important context for understanding proteins functions. In $\sim 60\%$ of protein-protein interactions the two interacting proteins share functional similarity, therefore identifying protein interactions is an important component of functional annotation. To overcome the gap between the known genes sequences and structures, we use GTD, a database that applies threading to predict the structure of all proteins with unknown structures. Furthermore, the protein-protein interaction prediction method builds on a number of relevant databases such as SCOPPI, PDB, GO, NetPro aiming to produce a comprehensive protein-protein interaction map staring from a set gene expression data. The key idea is using a sophisticated way of integrating this databased to help produce a high quality interaction network that helps uncovering some of the unknown behaviours of biological systems. The main contribution of this work is the use of interaction inference from known structures to predict novel protein-protein structures. First, the Genomic Threading Database (GTD) as fold recognition method to assign SCOP structural families to the proteins in our data sets is used. For the assigned SCOP domains, SCOPPI is used to identify interacting domain pairs. In this step, two proteins are considered as interacting if each contains a domain where there is structural evidence for such a domain–domain interaction according to SCOPPI. The evidence interaction then serves as a structural template to model the predicted interaction. For the resulting predicted interactions we further perform a refinement of predicted protein-protein interactions using interface conservation evaluation, considering only hetero-inter interactions and using protein localisation to filter out proteins in different cellular locations(GO). As a complimentary step we apply a pathway analysis approach, where the aim is to construct a cancer related pathway network that resembles the regulatory circuits which are disrupted in the cell.

Although the method can be used to analyse a wide range of high throughput data, it was mainly developed to analyse a set of PDAC gene expression data. The data set was obtained by integrating various analyses of the gene expression profiles of PDAC from Affymetrix GeneChip experiments such as microdissection, systematic isolation of genes and the meta-analysis of PDAC gene expression profiles from publicly available data

Taking into consideration the different aspects of how cancer behaves the method revealed interesting results that mainly short-list a huge number of genes into more interesting and relevant ones for cancer researchers where they can be tested in the lab. The results bellow are examples where our work was able to provide a positive answer to the open question number one that protein-protein interaction data indeed enhance knowledge discovery from gene expression data analysis to reveal potential markers and therapeutic targets for Pancreatic cancer. The protein-protein prediction method was programmed using Python.

1. A novel pancreatic cancer network of known and predicted protein-protein interactions By linking the pathway approach, known interactions and structure-based interaction predictions, we produce a detailed PDAC cell map. The map highlighted some clusters of interacting proteins that were further investigated. An interesting example is for a predicted interaction between transmembrane protease, serine 4 (TMPRSS4) and tissue factor pathway inhibitor 2 (TFPI2). We hypothesise that TFPI2 acts as a natural inhibitor of TMPRSS4. Since TFPI2 is downregulated, the upregulated TMPRSS4 is no longer inhibited and might facilitate tissue invasion. Another example that was revealed by this analysis is an interesting interaction between CDKN3 and CDC2L1. The interactions may prove valuable to improve our understanding of the regulatory mechanisms underlying the development of pancreatic cancer. Furthermore, our results indicate that in pancreatic cancer the calcium signalling pathway is affected.

2. Potential therapeutic targets for the treatment of PDAC using a novel drug (BVDU) We performed docking experiments using a docking technique based on Based on Shape Complementarity Principles. The results indicate that BVDU is able to bind to the active site of TMPRSS4, KIF20A and NNMT. The three genes were among the predicted protein-protein interactions of the proteins that changed expression after treatment with BVDU alone or in combination (Gemcitabine/BVDU). Although small molecule docking might produce some false positives, such analysis provide potential candidates that still needs to be tested experimentally. The docking results is an indication that these proteins may play a role as targets of BVDU.

3. Highlights the importance of the Co-expression of KLK6 and KLK10 as prognostic factor for survival in pancreatic ductal adenocarcinoma KLK10 and KLK6 are among the most highly and specifically overexpressed genes in pancreatic cancer compared with normal and benign pancreas tissues. The network of known and predicted interactions of KLK10 and KLK6 implies that it might have a role in the pathophysiology of PDAC.

4. Pancreatic cancer related apoptosis pathway We used our interaction prediction method to aid the construction of a dynamic map of apoptosis where the predicted interactions cans shed light on alternative ways of the effect of simultaneous gene silencing to examine its effects on apoptosis pathway.

6.1.2 Open problem 2: Does reasoning over molecular networks facilitate the analysis of gene expression data?

In chapter 4 we were aiming to provide an answer to the second open question. We developed BioRevise a belief revision that uses a high level representation of inhibition of enzyme-catalysed reactions, to reason over metabolic networks. We evaluated the BioRevise system with two cases. First a metabolic disorder example where BioRevise successfully identified the inhibition of the enzyme glucose-6-phosphatase (EC:3.1.3.9) as responsible for the Glycogen storage disease type I, which according to literature is known to be the main reason for this disease. The second example is the effect of the PDAC deregulated genes on the pancreas associated pathway “Glycolysis / Gluconeogenesis”. We could show that gene expression data from cancer tissues can be interpreted in terms of the metabolic pathways in which some of the co-regulated genes are involved in. For the knowledge modelling code of the Biorevise system we used extended logic program formalism and the user interface was coded in Java.

Acknowledgement

I would like to express my thanks and gratitude to all those who gave me the possibility to complete this thesis. I am deeply indebted to my thesis supervisor Prof. Michael Schroeder for guiding me through this research, with good advice, stimulating suggestions and encouragement that helped me in all the time of research and writing of this thesis. I would also like to thank Dr. Christian Pilarsky and his group from the Department of Visceral-, Thoracic- and Vascular Surgery, University Hospital, Dresden, and Dr. Jörg-Christian Heinrich from RESprotect GmbH, Dresden, for their collaboration and for allowing me to use their experimental data.

I would also like to extend my thanks to the Bioinformatics group members in general and specially those whom i had direct collaboration with during my work. I am particularly grateful to Dr. Andreas Henschel for reading the manuscript at least ten times and Dr. Christof Winter for always providing biological and medical information that helped me through my research and for spreading the good spirit when things do not go as planned.

I also thank my friends for all their help, support, interest and valuable hints.

A special thought for my family who gave me unconditional love, encouragement and support, specially my sister Randa for always being there for me and for her valuable advice of how to handle tough times.

Bibliography

S. Aerts, D. Lambrechts, S. Maity, L. Peter Van Loo, B. Coessens, S. Frederik De Smet, LC. Tranchevent, M. Bart De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–44, 2006.

A. J. Aguirre, C. Brennan, G. Bailey, R. Sinha, B. Feng, C. Leo, Y. Zhang, J. Zhang, J. D. Gans, N. Bardeesy, C. Cauwels, C. Cordon-Cardo, M. S. Redston, R. A. DePinho, and L. Chin. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, 101(24):9067–9072, 2004.

B. Alberts, D. Bray, A. Jonhson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology: an introduction to the molecular biology of the cell*, chapter 3 Energy, Catalysis, and Biosynthesis, pages 77–106. Garland Publishing, Inc., 1998.

P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of National Academy of Sciences USA*, 99(9):5896–5901, 2002.

P. Aloy and R. B. Russell. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197, 2006.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipmanl. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

P. Ariapart, S. Bergstedt-Lindqvist, H. Vanessa van Harmelen, J. Permert, F. Wang, and I. Lundkvist. Resection of pancreatic cancer normalizes the preoperative increase of tumor necrosis factor alpha gene expression. *Pancreatology*, 2(5):491–4, 2002.

M. Ashburner, C. A. Ball, J. A. Blake, and D. Botstein. Gene ontology: tool for the unification of biology. *nature genetics*, 25(1):25–29, 2000.

A. S. Aytuna, A. Gursoy, and O. Keskin. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855, 2005.

M. Bada, R. Stevens, C. Goble, Y. Gil, M. Ashburner, J. A. Blake, J. M. Cherry, M. Harris, and S. Lewis. A short study on the success of the gene ontology, 2004.

G. D. Bader and C. W. Hogue. Bind–a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, 16(5):465–77, 2000.

A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. The universal protein resource (uniprot). *Nucleic Acids Res*, 33(Database issue):D154–9, 2004.

T. Beissbarth and T.P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–5, 2004.

A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(Suppl 1):i38–46, 2005.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

JL. Blanchard, DL. Bulmore, AD. Farmer, M. Gonzales, PA. Steadman, ME. Waugh, ST. Wlodek, and P. Mendes. *Pathdb: a second generation metabolic database.* Stellenbosch University Press, 2000.

Michael L Blinov, James R Faeder, Byron Goldstein, and William S Hlavacek. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

A. Bockmayr and A. Courtois. Using hybrid concurrent constraint programming to model dynamic biological systems, 2002.

A. J. Bordner and R. Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60(3):353–366, 2005. doi: 10.1002/prot.20433.

C. A. Borgono and E. P. Diamandis. The emerging roles of human tissue kallikreins in cancer. *Nat Rev Cancer*, 4(11):876–90, 2004.

T.L. Bowles, R. Kim, J. Galante, C.M. Parsons, S. Virudachalam, H. Kung, and R.J. Bold. Pancreatic cancer cell lines deficient in argininosuccinate synthetase are sensitive to arginine deprivation by arginine deiminase. *Int J Cancer*, 123(8):1950–5, 2008.

J. R. Bradford and D. R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94, 2005.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C.P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nature Genetics*, 29: 365–371, 2001.

JM. Brown and LD. Attardi. The role of apoptosis in cancer development and treatment response. *Nat Rev Cancer*, 5(3):231–7, 2005.

L. Calzone, N. Chabrier-Rivier, F. Fages, and S. Soliman. A machine learning approach to biochemical reaction rules discovery. In *Proceedings of FOSBE'05)*, 2005.

E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:262–6, 2004.

D. Cao, S. R. Hustinx, G. Sui, P. Bala, N. Sato, S. Martin, A. Maitra, K. M. Murphy, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, A. Pandey, and R. H. Hruban. Identification of novel highly expressed genes in pancreatic ductal adenocarcinomas through a bioinformatics analysis of expressed sequence tags. *Cancer Biol Ther*, 3(11):1081–1089, 2004.

R. Caspi, H. Foerster, C.A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S.Y. Rhee, C. Tissier, P. Zhang, and P.K. Karp. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acid Research*, 34(Database issue): D511–6, 2006.

K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak, and J. J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell*, 15(8):3841–62, 2004.

J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. mbrough, M. S. Cline, G. Liu, R. Shigeta, and D. K. S. Rose. Netaffx gene ontology mining tool: A visual approach for microarray data analysis. *Bioinformatics*, 20(9):1462–3, 2004.

K. Cheung, J. Hager, D. Pan, R. Srivastava, S. Mane, Y. Li, P. Miller1, and K. R. Williams. Karma: a web server application for comparing and annotating heterogeneous microarray platforms. *Nucleic Acids Research*, 32:W441–4, 2004.

Timothy S. C. Chou and Marianne Winslett. The implementation of a model-based belief revision system. *SIGART Bull.*, 2(3):28–34, 1991. ISSN 0163-5719. doi: http://doi.acm.org/10.1145/122296.122301.

H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

M. Coleman, S. Walsh, L. Li, M. Hinkhouse, D. Spitz, and J. Cullen. Inhibition of glucose metabolism in pancreatic cancer induces cytotoxicty via metabolic oxidative stress. *Journal of the American College of Surgeons*, 199,(3):24, 2004.

D. Coppola. Molecular prognostic markers in pancreatic cancer. *Cancer Control*, 7(5):421–427, 2000.

J. Cruz and P. Barahona. Constraint reasoning in deep biomedical models. *Artif Intell Med*, 34 (1):77–88, 2005.

J M Culouscou and M Shoyab. Purification of a colon cancer cell growth inhibitor and its identification as an insulin-like growth factor binding protein. *Cancer Res*, 51(11):2813–9, 1991.

C. V. Damásio, L. M. Pereira, and M. Schroeder. REVISE: Logic programming and diagnosis. In Jürgen Dix, Ulrich Furbach, and Anil Nerode, editors, *Proocedings of the Fourth International Conference on Logic Programming and Non-Monotonic Reasoning*, Lecture Notes in Artificial Intelligence, pages 353–362. Springer-Verlag, 1997.

F. P. Davis and A. Sali. Pibase: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–7, 2005.

G. Dawelbait, C. Pilarsky, Y. Zhang, R. Grützmann, and M. Schroeder. Structural protein interactions predict kinase-inhibitor interactions in upregulated pancreas tumour genes expression data. In *CompLife*, pages 1–11, 2005.

C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5): 349–56, 2002.

A. Degterev, M. Boyce, and J. Yuan. A decade of caspases. *Oncogene*, 22(53):8543–67, 2003.

A. Deshpande, P. Sicinski, and P. W. Hinds. Cyclins and cdks in development and cancer: a perspective. *Oncogene*, 24(17):2909–2915, 2005.

Y. Deville, D. Gilbert, J. Van Helden, and W. Shoshana. An overview of data models for the analysis of biochemical pathways. *Briefings in Bioinformatics*, 4:246–259, 2003.

A. Djebbari and J. Quackenbush. Seeded bayesian networks: constructing genetic networks from microarray data. *BMC Syst Biol*, 2(1):57, 2008.

P. Dönnes, A. Höglund, M. Sturm, N. Comtesse, E. Backe, C.and Meese, O. Kohlbacher, and H.P. Lenhof. Integrative analysis of cancer-related data using cap. *FASEB J*, 18(12):1465–7, 2004.

M. Ducreux, V. Boige, and D. Malka. Emerging drugs in pancreatic cancer. *Expert Opin Emerg Drugs*, 9(1):73–89, 2004.

M. S. Duxbury, H. Ito, E. Benoit, M. J. Zinner, S. W. Ashley, and E. E. Whang. Rna interference targeting focal adhesion kinase enhances pancreatic adenocarcinoma gemcitabine chemosensitivity. *Biochem Biophys Res Commun*, 311(3):786–92, 2003.

S. Eker, M. Knapp, K. Laderoute, P. Lincoln, and a. C. Talcott. Pathway logic: Executable models of biological networks. In *Fourth International Workshop on Rewriting Logic and Its Applications (WRLA'2002*, volume 71 of *Electronic Notes in Theoretical Computer Science*. Elsevier, 2002.

A. H. Elcock and J. A. McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A*, 98(6):2990–2994, 2001.

M. Erkan, J. Kleeff, I. Esposito, T. Giese, K. Ketterer, M. W. Büchler, N. A. Giese, and H. Friess. Loss of bnip3 expression is a late event in pancreatic cancer contributing to chemoresistance and worsened prognosis. *Oncogene*, 24(27):4421–32, 2005.

J. Espadaler, O. Romero-Isart, R. M. Jackson, and B. Oliva. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, 21(16):3360–3368, 2005.

F. Fages, S. Soliman, and N. Chabrier-Rivier. Modelling and querying interaction networks in the biochemical abstract machine biocham. *Journal of Biological Physics and Chemistry*, 4 (2):64–73, 2005.

R. Fahrig, J. C. Heinrich, B. Nickel, F. Wilfert, C. Leisser, G. Krupitza, C. Praha, D. Sonntag, B. Fiedler, H. Scherthan, and H. Ernst. Inhibition of induced chemoresistance by cotreatment with (e)-5-(2-bromovinyl)-2'-deoxyuridine (rp101). *Cancer Res*, 63(18):5745–53, 2003.

R. Fahrig, D. Quietzsch, J. C. Heinrich, V. Heinemann, S. Boeck, R. M. Schmid, C. Praha, A. Liebert, D. Sonntag, G. Krupitza, and M. Hnel. Rp101 improves the efficacy of chemotherapy in pancreas carcinoma cell lines and pancreatic cancer patients. *Anticancer Drugs*, 17(9):1045–56, 2006.

S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6, 1989.

R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 2007.

J. B. Fleming, G. Shen, S. E. Holloway, M. Davis, and R. A. Brekken. Molecular consequences of silencing mutant k-ras in pancreatic cancer cells: justification for k-ras-directed therapy. *Mol Cancer Res*, 3(7):413–423, 2005.

J Freudenberg and P Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2:S110–5, 2002.

S Fulda and K-M Debatin. Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapy. *Oncogene*, 25(34):4798–811, 2006.

P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–83, 2004.

M. Y. Galperin and E. V. Koonin. Who's your neighbor? new computational approaches for functional genomics. *Nat Biotechnol*, 18(6):609–13, 2000.

A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

C. Gennatas, V. Michalaki, D. Mouratidou, N. Tsavaris, C. Andreadis, A. Photopoulos, and D. Voros. Gemcitabine combined with 5-fluorouracil for the treatment of advanced carcinoma of the pancreas. *In Vivo*, 20(2):301–5, 2006.

M. C. Ghosh, L. Grass, A. Soosaipillai, G. Sotiropoulou, and E. P. Diamandis. Human kallikrein 6 degrades extracellular matrix proteins and may enhance the metastatic potential of tumour cells. *Tumour Biol*, 25(4):193–9, 2004.

L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. J. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

L. Glass and S. A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*, 39(1):103–29, 1973.

A L Glazyrin, V N Adsay, V K Vaitkevicius, and F H Sarkar. Cd95-related apoptotic machinery is functional in pancreatic cancer cells. *Pancreas*, 22(4):357–65, 2001.

K. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabasi. The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685–90, 2007.

S. Gong, G. Yoon, I. Jang, D. Bolser, P. Dafas, M. Schroeder, H. Choi, Y. Cho, K. Han, S. Lee, H. Choi, M. Lappe, L. Holm, S. Kim, D. Oh, and J. Bhak. Psibase: a database of protein structural interactome map (psimap). *Bioinformatics*, 21(10):2541–2543, 2005.

B. C. Goodwin. Temporal organization in cells : a dynamic theory of cellular control processes. *Academic Press, New York*, 4, 1963.

R. Grützmann, M. Foerder, I. Alldinger, E. Staub, T. Brümmendorf, S. Rpcke, X. Li, G. Kristiansen, R. Jesnowski, B. Sipos, M. Lhr, J. Lüttges, D. Ockert, G. Klppel, H. D. Saeger, and C. Pilarsky. Gene expression profiles of microdissected pancreatic ductal adenocarcinoma. *Virchows Arch*, 443(4):508–17, 2003.

R. Grutzmann, C. Pilarsky, E. Staub, A. O. Schmitt, M. Foerder, T. Specht, B. Hinzmann, E. Dahl, I. Alldinger, A. Rosenthal, D. Ockert, and H. Saeger. Systematic isolation of genes differentially expressed in normal and cancerous tissue of the pancreas. *Pancreatology*, 3(2): 169–78, 2003.

R. Grutzmann, C. Pilarsky, O. Ammerpohl, J. Luttges, A. Bohme, B. Sipos, M. Foerder, I. Alldinger, B. Jahnke, H. K. Schackert, H. Kalthoff, B. Kremer, G. Kloppel, and H. D. Saeger. Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays. *Neoplasia*, 6(5):611–22, 2004a.

R. Grutzmann, H. D. Saeger, J. Luttges, H. Schackert, Kalthoff H., G. Kloppel, and C. Pilarsky. Microarray-based gene expression profiling in pancreatic ductal carcinoma: status quo and perspectives. *International Journal of Colorectal Disease*, 19:401– 413, 2004b.

R. Grutzmann, H. Boriss, O. Ammerpohl, J. Luttges, H. Kalthoff, H. K. Schackert, G. Kloppel, H. D. Saeger, and C. Pilarsky. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 24(32):5079–88, 2005.

A. Guffanti. Modeling molecular networks: a systems biology approach to gene function. *Genome Biol*, 3(10):40–31, 2002.

Anna S Gukovskaya and Stephen J Pandol. Cell death pathways in pancreatitis and pancreatic cancer. *Pancreatology*, 4(6):567–86, 2004.

R. Hamacher, R.M. Schmid, D. Saur, and G. Schneider. Apoptotic pathways in pancreatic ductal adenocarcinoma. *Mol Cancer*, 7:64, 2008.

D. Han, H. Kim, W. Jang, S. Lee, and J. Suh. Prespi: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res*, 32(21):6312–6320, 2004.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

G J Hannon, D Casso, and D Beach. Kap: a dual specificity phosphatase that interacts with cyclin-dependent kinases. *Proc Natl Acad Sci U S A*, 91(5):1731–5, Mar 1994.

M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz la Cruz, P Tonellato, P Jaiswal, T Seigfried, and R White. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32:258–261, 2004.

S. Hennig, D. Groth, and H. Lehrach. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acid Research*, 31(13):3712–3715, 2003a.

Steffen Hennig, Detlef Groth, and Hans Lehrach. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Res*, 31(13):3712–5, 2003b.

H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):452–455, 2004.

A. F. Hezel, A. C. Kimmelman, B. Z. Stanger, N. Bardeesy, and R. A. Depinho. Genetics and biology of pancreatic ductal adenocarcinoma. *Genes Dev*, 20(10):1218–49, 2006a.

A. F. Hezel, A. C. Kimmelman, B. Z. Stanger, N. Bardeesy, and R. A. Depinho. Genetics and biology of pancreatic ductal adenocarcinoma. *Genes Dev*, 20(10):1218–49, 2006b.

M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res*, 35(Database issue):D721–6, 2007.

Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

R. Hofestadt and S. Thelen. Quantitative modeling of biochemical networks. *In Silico Biol*, 1 (1):39–53, 1998.

R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat Genet*, 36(7): 664, 2004.

B. Huang and M. Schroeder. Using residue propensities and tightness of fit to improve rigid-body protein-protein docking. *In: Proceedings of the German Conference on Bioinformatics. GI LNI71*, P-71:159–173, 2005. ISSN 1362-4962.

S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res*, 2008.

Jochen Hurlebaus, Arne Buchholz, Wolfgang Alt, Wolfgang Wiechert, and Ralf Takors1. Mmt - a pathway modeling tool for data from rapid sampling experiments. *In Silico Biology*, 2002.

M. Huss and P. Holme. Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks, 2006.

S. R. Hustinx, D. Cao, A. Maitra, N. Sato, S. T. Martin, D. Sudhir, C. Iacobuzio-Donahue, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, J. Mollenhauer, A. Pandey, and R. H. Hruban. Differentially expressed genes in pancreatic ductal adenocarcinomas identified through serial analysis of gene expression. *Cancer Biol Ther*, 3(12):1254–1261, 2004.

T. R. Hvidsten, A. Lgreid, and J. Komorowski. Learning rule-based models of biological process from gene expression time profiles using gene ontology. *bioinformatics*, 19(9):1116–1123, 2003.

C. A. Iacobuzio-Donahue, A. Maitra, G. L. Shen-Ong, T. van Heek, R. Ashfaq, R. Meyer, K. Walter, K. Berg, M. A. Hollingsworth, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, and R. H. Hruban. Discovery of novel tumor markers of pancreatic cancer using global gene expression technology. *Am J Pathol*, 160(4):1239–1249, 2002.

C. A. Iacobuzio-Donahue, R. Ashfaq, A. Maitra, N. V. Adsay, G. L. Shen-Ong, K. Berg, M. A. Hollingsworth, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, and R. H. Hruban. Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Res*, 63(24):8614–22, 2003a.

C. A. Iacobuzio-Donahue, R. Ashfaq, A. Maitra, N. V. Adsay, G. L. Shen-Ong, K. Berg, M. A. Hollingsworth, J. L. Cameron, C. J. Yeo, S. E. Kern, M. Goggins, and R. H. Hruban. Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Res*, 63(24):8614–8622, 2003b.

R. Ishizawar and S. J. Parsons. c-src and cooperating partners in human cancer. *Cancer Cell*, 6 (3):209–214, 2004.

T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of National Academy of Sciences USA*, 98(8):4569–4574, 2001.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, 2003.

S. Jones, X. Zhang, D.W. Parsons, J.C. Lin, R.J. Leary, P. Angenendt, H. Mankoo, P. andCarter, H. Kamiyama, A. Jimeno, S. Hong, B. Fu, M. Lin, E.S Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D.R. Smith, M. Hidalgo, S.D. Leach, A.P. Klein, E.M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J.R. Eshleman, S.E. Kern, R.H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V.E. Velculescu, and K.W. Kinzler. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801–6, 2008.

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. B. de, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–32, 2005.

M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res.*, 30(1):42–46, 2002.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32:D277–D280, 2004.

M. Kanehisa, 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:D354–D357, 2006.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 2007.

G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 2008.

P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrn, S. Tsoka, N. Darzentas, V. Kunin, and N. Lpez-Bigas. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, 33(19):6083–9, 2005.

M. Katoh and M. Katoh. Human fox gene family (review). *Int J Oncol*, 25(5):1495–1500, 2004.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, 22(3):437–67, 1969.

H. Kayed, J. Kleeff, T. Osman, S. Keleg, M. W. Buchler, and H. Friess. Hedgehog signaling in the normal and diseased pancreas. *Pancreas*, 32(2):119–129, 2006.

E. Kebebew, M. Peng, E. Reiff, Q. Y. Duh, O. H. Clark, and A. McMillan. Ecm1 and tmprss4 are diagnostic markers of malignant thyroid neoplasms and improve the accuracy of fine needle aspiration biopsy. *Ann Surg*, 242(3):353–61; discussion 361–3, 2005.

S. P. Khosravi and M. d. l. E. V. Daz. [pancreatic adenocarcinoma: therapeutical update]. *An Med Interna*, 22(8):390–4, 2005.

J. Kim and J. C. Park. *Annotation of Gene Products in the Literature with Gene Ontology Terms using Syntactic Dependencies*. Springer Berlin / Heidelberg, 2005.

W. K. Kim, D. M. Bolser, and J. H. Park. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, 20(7):1138–1150, 2004.

Wan Kyu Kim, Andreas Henschel, Christof Winter, and Michael Schroeder. The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Computational Biology*, 2(9):e124, 2006.

A. Koike and T. Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel*, 17(2):165–173, 2004.

F. Kolpakov, V. Poroikov, R. Sharipov, Y. Kondrakhin, A. Zakharov, A. Lagunin, L. Milanesi, and A. Kel. CYCLONET–an integrated database on cell cycle regulation and carcinogenesis. *Nucleic Acids Res*, 35(Database issue):D550–6, 2007.

A. Kumar and B. Smith. On controlled vocabularies in biology : A case study in the gene ontology. *Drug Discovery Today*, 6(5):246–252, 2004.

A. Kumar and B. Smith. A framework for protein classification. In *Proceedings of the German Confrence on Bioinformatics*, 2003.

A. Kumar, B. Smith, and C. Borgelt. Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 31–38, Geneva, Switzerland, 2004.

Evelina Lamma, Lu'is Moniz Pereira, and Fabrizio Riguzzi. Belief revision by multi-agent genetic search. In *In In Proc. of the 2nd International Workshop on Computational Logic for Multi-Agent Systems, Paphos*, 2001.

I. V. Lebedeva, D. Sarkar, Z. Su, R. V. Gopalkrishnan, M. Athar, A. Randolph, K. Valerie, P. Dent, and P. B. Fisher. Molecular target-based therapy of pancreatic cancer. *Cancer Res*, 766(4):72403–2413, 2006.

I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork. Smart 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 34(Database issue):D257–60, 2006.

S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. Vidalain, J. J. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

S Li, P Brazhnik, B Sobral, and JJ Tyson. A quantitative study of the division cycle of caulobacter crescentus stalked cells. *PLoS Comput Biol*, 4(1):e9, 2008.

G. Lolli, E. D. Lowe, N. R. Brown, and L. N. Johnson. The crystal structure of human CDK7 and its protein recognition properties. *Structure*, 12(11):2067–79, 2004.

Rita Bayer Lopes, Rathi Gangeswaran, Iain A McNeish, Yaohe Wang, and Nick R Lemoine. Expression of the iap protein family is dysregulated in pancreatic cancer cells and is important for resistance to chemotherapy. *Int J Cancer*, 120(11):2344–52, 2007.

P. Lord, R. Stevens, A. Bras, and C. A. Goble. Investigating semantic similarity measures across the gene ontology:the relationship between sequence and annotaion. *Bioinformatics*, 2002.

S. Maere, K. Heymans, and M. Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9, 2005.

D. Marc A van Driel, J. Bruggeman, G. Vriend, H.G. Brunner, and J.A.M. Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–42, 2006.

E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285 (5428):751–753, 1999.

D. M. Martin, M. Berriman, and G. J. Barton. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC BIOINFORMATICS*, 2004.

M. A. Mart-Renom, A. C. Stuart, A. Fiser, R. Snchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29: 291–325, 2000. ISSN 1056-8700. doi: 10.1146/annurev.biophys.29.1.291.

L. J. McGuffin and D. T. Jones. Improvement of the genthreader method for genomic fold recognition. *Bioinformatics*, 19(7):874–881, 2003a.

L. J. McGuffin and D. T. Jones. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 19(7):874–881, 2003b.

P. Mehlen and A. Puisieux. Metastasis: a question of life or death. *Nat Rev Cancer*, 6(6): 449–58, 2006.

S. Menard, P. Casalini, M. Campiglio, S. M. Pupa, and E. Tagliabue. Role of HER2/neu in tumor progression and therapy. *Cell Mol Life Sci*, 61(23):2965–78, 2004.

C. V. Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–7, 2005.

K. D. Mertz, S. R. Setlur, S. M. Dhanasekaran, F. Demichelis, S. Perner, S. Tomlins, J. Tchinda, B. Laxman, R. L. Vessella, R. Beroukhim, C. Lee, A. M. Chinnaiyan, and M. A. Rubin. Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model. *Neoplasia*, 9(3):200–6, 2007.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536, 1995.

H. Nagahara, K. Mimori, T. Utsunomiya, G. F. Barnard, M. Ohira, K. Hirakawa, and M. Mori. Clinicopathologic and biological significance of kallikrein 6 overexpression in human gastric cancer. *Clin Cancer Res*, 11(19 Pt 1):6800–6, 2005.

V. Neduva and R. B. Russell. Dilimot: discovery of linear motifs in proteins. *Nucleic Acids Res*, 34(Web Server issue):350–355, 2006. doi: 10.1093/nar/gkl159.

Wolfgang Nejdl and Brigitte Giefer. Drum:reasoning without conflicts and justifications. In *In 5th International Workshop on Principles of Diagnosis (DX-94*, pages 226–233, 1994.

A. Ng, B. Bursteinas, Q. Gao, E. Mollison, and M. Zvelebil. pSTIING: a 'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. *Nucleic Acids Res*, 34(Database issue):D527–34, 2006.

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27(1):29–34, 1999.

U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy. Prism: protein interactions by structural matching. *Nucleic Acids Res*, 33(Web Server issue):331–336, 2005.

K. Ozaki, M. Nagata, M. Suzuki, T. Fujiwara, Y. Miyoshi, O. Ishikawa, H. Ohigashi, S. Imaoka, E. Takahashi, and Y. Nakamura. Isolation and characterization of a novel human pancreas-specific gene, pancpin, that is down-regulated in pancreatic cancer cells. *Genes Chromosomes Cancer*, 22(3):179–85, 1998.

Y. W. Park, S. Kim, K. JO, J. H. Park, H Jung, K. P. Lee, S. J. Park, Y Jang, S Choi, J. G. Jung, H. J. Hong, J. H. Yoon, and J. H. Park. An anticancer drug comprising inhibitor of tmprss4, 2007.

M. Pellegrini, EM. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, 1999.

S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. G. K, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. C. R, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. C. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, 2003.

C.D. Petraki, V.N. Karavana, K.I. Revelos, Y. Luo, and E.P. Diamandis. Immunohistochemical localization of human kallikreins 6 and 10 in pancreatic islets. *Histochem J*, 34(6-7):313–22, 2002 Jun-Jul.

C. Petri. Kommunikation mit automaten. *Schriften des Institutes fur Instumentelle mathematik*, 1962.

J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*, 26 (16):1781–802, 2005.

C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser. Alibaba: Pubmed as a graph. *Bioinformatics*, 22(19):2444–5, 2006.

Y. Pommier, O. Sordet, S. Antony, R. L. Hayward, and K. W. Kohn. Apoptosis defects and chemotherapy resistance: molecular interaction maps and networks. *Oncogene*, 23(16): 2934–49, 2004.

H. T. Poon RY. Dephosphorylation of cdk2 thr160 by the cyclin-dependent kinase-interacting phosphatase kap in the absence of cyclin. *Science*, 1995.

P. Pospisil, L. K. Iyer, S. J. Adelstein, and A. I. Kassis. A combined approach to data mining of textual and structured data to identify cancer-related targets. *BMC Bioinformatics*, 7:354, 2006.

O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Sraphin. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–29, 2001.

P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. A. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson. Elm server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, 31(13):3625–3630, 2003.

J. Quackenbush. Extracting biology from high-dimensional biological data. *J Exp Biol*, 210(Pt 9):1507–17, 2007.

J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of Helicobacter pylori. *Nature*, 409(6817):211–215, 2001.

C. N. Rao, P. Reddy, Y. Liu, E. O'Toole, D. Reeder, D. C. Foster, W. Kisiel, and D. T. Woodley. Extracellular matrix-associated serine protease inhibitors (mr 33,000, 31,000, and 27,000) are single-gene products with differential glycosylation: cdna cloning of the 33-kda inhibitor reveals its identity to tissue factor pathway inhibitor-2. *Arch Biochem Biophys*, 335(1):82–92, 1996.

A. Regev, E. M. Panina, W. Silverman, L. Cardelli, and E. Shapiro. Bioambients: An abstraction for biological compartments. *Theoretical Computer Science*, pages 141–167, 2004.

D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–9, 2005.

C. Rosa and B. Smith. The role of foundational relations in the alignment of biomedical ontologies. *MEDIINFO*, 2004.

J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

F. Rückert, M. Hennig, C. D. Petraki, D. Wehrum, M. Distler, A. Denz, M. Schrder, G. Dawelbait, H. Kalthoff, H. D. Saeger, E. P. Diamandis, C. Pilarsky, and R. Grützmann. Co-expression of klk6 and klk10 as prognostic factors for survival in pancreatic ductal adenocarcinoma. *Br J Cancer*, 2008.

F. Sahin, A. Maitra, P. Argani, N. Sato, N. Maehara, E. Montgomery, M. Goggins, R.H. Hruban, and G.H. Su. Loss of stk11/lkb1 expression in pancreatic and biliary neoplasms. *Mod Pathol*, 16(7):686–91, 2003.

P. Saraiya, C. North, and C. Duca. Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Information Visualization*, pages 1–15, 2005.

N. Sato, A. R. Parker, N. Fukushima, Y. Miyagi, C. A. Iacobuzio-Donahue, J. R. Eshleman, and M. Goggins. Epigenetic inactivation of tfpi-2 as a common mechanism associated with growth and invasion of pancreatic ductal adenocarcinoma. *Oncogene*, 24(5):850–8, 2005.

K Satoh, K Kaneko, M Hirota, A Masamune, A Satoh, and T Shimosegawa. Tumor necrosis factor-related apoptosis-inducing ligand and its receptor expression and the pathway of apoptosis in human pancreatic cancer. *Pancreas*, 23(3):251–8, 2001.

T. K. Sawyer. Novel oncogenic protein kinase inhibitors for cancer therapy. *Curr Med Chem Anti-Canc Agents*, 4(5):449–455, 2004.

D. Schneidman-Duhovny, I. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H.J. Wolfson. Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*, 52(1):107–12, 2003.

D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson. Patchdock and symmdock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 33(Web Server issue):W363–7, 2005.

S. Schulz, P. Daumke, B. Smith, and U. Hahn. How to distinguish parthood from location in bioontologies. *AMIA Annu Symp Proc*, pages 669–673, 2005.

H. Schulze-Bergkamen and P.H. Krammer. Apoptosis in cancer–implications for therapy. *Semin Oncol*, 31(1):90–119, 2004.

M. Smid and L. C. J. Dorssers. Go-mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms. *Bioinformatics*, 2004.

B. Smith. Beyond concepts: Ontology as reality representaion. In *Proceedings of the International Conference on Formal Ontology and information Systems*, 2004.

B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology, 2003.

B. Smith, J. Khler, and A. Kumar. On the application of formal principles to life science data: A case study in the gene ontology. *DILS*, 2004.

P. Sova, Q. Feng, G. Geiss, T. Wood, R. Strauss, V. Rudolf, A. Lieber, and N. Kiviat. Discovery of novel methylation biomarkers in cervical carcinoma by global demethylation and microarray analysis. *Cancer Epidemiol Biomarkers Prev*, 15(1):114–23, 2006.

E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692, 2001. doi: 10.1006/jmbi.2001.4920.

A. Stein, R. B. Russell, and P. Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, 33:D413–D417, 2005.

J. Sun, J. Xu, Z. Liu, Q. Liu, A. Zhao, T. Shi, and Y. Li. Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, 21(16):3409–3415, 2005.

A. Tamaddoni-Nezhad, A. C. Kakas, S. Muggleton, and F. Pazos. Modelling inhibition in metabolic pathways through abduction and induction. In *ILP*, volume 3194, pages 305–322, 2004.

K. Taniuchi, H. Nakagawa, T. Nakamura, H. Eguchi, H. Ohigashi, O. Ishikawa, T. Katagiri, and Y. Nakamura. Down-regulation of RAB6KIFL/KIF20A, a kinesin involved with membrane trafficking of discs large homologue 5, can attenuate growth of pancreatic cancer cell. *Cancer Res*, 65(1):105–12, 2005.

S. P. Thayer, M. P. di Magliano, P. W. Heiser, C. M. Nielsen, D. J. Roberts, G. Y. Lauwers, Y. P. Qi, S. Gysin, C. Fernandez-del Castillo, V. Yajnik, B. Antoniu, M. McMahon, A. L. Warshaw, and M. Hebrok. Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. *Nature*, 425(6960):851–856, 2003.

R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–85, 1973.

R. Tibes, J. Trent, and R. Kurzrock. Tyrosine kinase inhibitors and the dawn of molecular cancer therapeutics. *Annu Rev Pharmacol Toxicol*, 45:357–384, 2005.

G. J. Tope, M. Gillespie, I. Vastrik, P. D Eustachio, E. Schmidt, B. d. Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein1. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33:D428–D432, 2005.

A Trauzold, S Schmiedel, C Rder, C Tams, M Christgen, S Oestern, A Arlt, S Westphal, M Kapischke, H Ungefroren, and H Kalthoff. Multiple and synergistic deregulations of apoptosis-controlling genes in pancreatic carcinoma cells. *Br J Cancer*, 89(9):1714–21, 2003.

P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–7, 2000.

W. S. Valdar and J. M. Thornton. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–124, 2001.

A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–73, 2002.

M. Veit and S. Herrmann. Model-view-controller and Object Teams: A perfect match of paradigms. In *AOSD'03*, pages 140–149, 2003.

Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–99, 2004.

M. Vogler, K. Dürr, M. Jovanovic, K. M. Debatin, and S. Fulda. Regulation of trail-induced apoptosis by xiap in pancreatic carcinoma cells. *Oncogene*, aop(current).

S. Volinia, R. Evangelisti, F. Francioso, D. Arcelli, M. Carella, and P. Gasparini. Goal: automated gene ontology analysis of expression profiles. *Nucleic Acids Research*, 24(1):W492–9, 2004.

C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–62, 2007.

C. Wallrapp, S. Hahnel, F. Muller-Pillasch, B. Burghardt, T. Iwamura, M. Ruthenburger, M. M. Lerch, G. Adler, and T. M. Gress. A novel transmembrane serine protease (tmprss3) overexpressed in pancreatic cancer. *Cancer Res*, 60(10):2602–6, 2000.

S. Westphal and H. Kalthoff. Apoptosis: targets in pancreatic cancer. *Mol Cancer*, 2:6, 2003.

JA. Williams. Intracellular signaling mechanisms activated by cholecystokinin-regulating synthesis and secretion of digestive enzymes in pancreatic acinar cells. *Annu Rev Physiol*, 63: 77–97, 2001.

C. Winter, T. Baust, B. Hoflack, , and M. Schroeder. A novel, comprehensive method to detect and predict protein-protein interactions applied to the study of vesicular trafficking. In *Proceedings of German Bioinformatics Conference GCB*, 2006a.

C. Winter, A. Henschel, W. K. Kim, and M. Schroeder. SCOPPI: A Structural Classification of Protein–Protein Interfaces. *Nucleic Acids Res*, 34(Database issue):310–314, 2006b.

I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, 2000.

J. Xu and L. Attisano. Mutations in the tumor suppressors Smad2 and Smad4 inactivate transforming growth factor beta signaling by targeting Smads to the ubiquitin-proteasome pathway. *Proc Natl Acad Sci U S A*, 97(9):4820–4825, 2000.

K. Yano, O. H. Petersen, and A. V. Tepikin. Computational models of calcium signaling in the pancreas- temporal and spatial regulations. *Genome Informatics*, 14:603–604, 2003.

G. M. Yousef, C. A. Borgoo, C. Popalis, G. M. Yacoub, M. E. Polymeris, A. Soosaipillai, and E. P. Diamandis. In-silico analysis of kallikrein gene expression in pancreatic and colon cancers. *Anticancer Res*, 24(1):43–51, 2004.

B. Zeeberg, W. Feng, G. Wang, M. Wang, A. Fojo, M. Sunshine, S. Narasimhan, D. Kane, W. Reinhold, S. Lababidi, K. Bussey, J. Riss, J. Barrett, and J. Weinstein. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4): R28, 2003.

A. Zhang. *Advanced analysis of gene expression microarray data*. World Scientific Publishing Co. Pte.Ltd, 2006.

B. Zhang, D. Schmoyer, S. Kirov1, and J. Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC BIOINFORMATICS*, 16(5), 2004.

Y. Zhang, I. Bhat, M. Zeng, G. Jayal, D. E. Wazer, H. Band, and V. Band. Human kallikrein 10, a predictive marker for breast cancer. *Biol Chem*, 387(6):715–21, 2006.

S. Zhong1, C. Li, and W. H. Wong. Chipinfo: software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Research*, 31(13):3483–3486, 2003.

M. Zhou and Y. Cu. Geneinfoviz: Constructing and visualizing gene relation networks. *In Silico Biology*, 2004.