Structural templates predict novel protein interactions and targets from pancreas tumour gene expression data

Gihan Dawelbait¹, Christof Winter¹, Yanju Zhang¹, Christian Pilarsky², Robert Grützmann², Jörg-Christian Heinrich³ and Michael Schroeder¹,*

¹Bioinformatics Group, Biotechnological Centre, TU Dresden, Dresden, Germany, ²Department of Visceral-, Thoracic- and Vascular Surgery, University Hospital, Dresden, Germany and ³RESprotect GmbH, Dresden, Germany

ABSTRACT

Motivation: Pancreatic ductal adenocarcinoma (PDAC) eludes early detection and is characterized by its aggressiveness and resistance to current therapies. 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.

Results: Here, we take such a network-centric approach to pancreas cancer by re-constructing networks from known interactions and by predicting novel protein interactions from structural templates. The pathways we find to be largely affected are signal transduction, actin cytoskeleton regulation, cell growth and cell communication.

Our analysis indicates that the alteration of the calcium pathway plays an important role in pancreas-specific tumorigenesis. Furthermore, our structural prediction method identifies 40 novel interactions including the tissue factor pathway inhibitor 2 (TFPI2) interacting with the transmembrane protease serine 4 (TMPRSS4). Since TMPRSS4 is involved in metastasis formation, we hypothesise that the upregulation of TMPRSS4 and the downregulation of its predicted inhibitor TFPI2 plays an important role in this process. Moreover, we examine the potential role of BVDU (RP101) as an inhibitor of TMPRSS4. BDVU is known to support apoptosis and prevent the acquisition of chemoresistance. Our results suggest that BVDU might bind to the active site of TMPRSS4, thus reducing its assistance in metastasis.

Contact: mfs@biotec.tu-dresden.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Biological background. The pancreas is located in the upper abdomen in close proximity to the duodenum. It serves two major functions: secretion of digestive enzymes by the pancreatic exocrine cells and production of hormones such as insulin, glucagon and somatostatin by the endocrine cells. Pancreatic cancer is the fourth leading cause of death due to cancer in virtually all industrialized countries. It accounts for 100,000 deaths per year in the United States and Europe. Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic neoplasm and is found in about 80% of pancreatic tumour cases (Hezel et al., 2006). Since pancreatic cancer is not only difficult to detect, but also difficult to treat, it has an extremely poor prognosis. To improve this prognosis, novel molecular markers for earlier diagnosis and targets for adjuvant or neoadjuvant treatment need to be identified.

Gene expression and pancreas cancer. Many researchers have carried out gene expression experiments coupled to computational analyses to identify relevant markers and targets for pancreas cancer (Aguirre et al., 2004; Cao et al., 2004; Grützmann et al., 2004a; Hezel et al., 2006; Hustinx et al., 2004; Iacobuzio-Donahue et al., 2002).

Using dimension reduction with principal identified seven genes involved in multiple cellular processes such as signal transduction (MIC-1), differentiation (DMBT1 and Neurgin), immune response (CD74), inflammation (CXCL2), cell cycle (CEB1) and enzymatic activity (Kalikirein 6). Pospisil et al., 2006 and Cao et al., 2004 developed promising approaches integrating a host of bioinformatics resources covering sequence and structural data. Cao’s analysis revealed the importance of CD29, INHBA, AKAP12, ELK3, FOXQ1, EIF5A2, and EFNA5, which were experimentally validated, while Pospisil et al. (2006) focused on alkaline phosphatase (various cancers), prostatic acid phosphatase, prostate-specific antigen (prostate cancer) and extracellular sulfatase 1 (pancreatic cancer).

A systems approach to pancreas cancer. 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 (Ashburner et al., 2000) and relevant protein interactions from Ingenuity’s Pathways Analysis. Recent databases such as pSTING (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 pancreas cancer such as a tyrosine kinase subnetwork including ERBB2, MUC1, SHC1, EPH2A and invasion signalling including NET1, RhoA, RhoC and RAC.

Over the past years, such a network-based approach has become possible. Fuelled by high-throughput interaction experiments (Gavin et al., 2002; Gavin et al., 2006; Giot et al., 2003; Ho et al., 2002; Ito et al., 2001; Li et al., 2003; Rain et al., 2001; Uetz et al., 2000) 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., 2004), HPRD (Peri et al., 2003), BIND (Bader et al., 2003), KEGG (Kanehisa et al., 2005), 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., 2006). Finally, there are many efforts to extract interactions from literature, among them iHOP (Hoffmann and Valencia, 2004) and ALI BABA (Plake et al., 2006).

These data repositories provide valuable resources for the prediction of protein–protein interactions. So far, sequence-based methods focused on gene context conservation (Galperin and Koonin, 2000), phylogenetic profiling (Pellegrini et al., 1999; Sun et al., 2005) and co-evolution of gene expression (Fraser et al., 2004). Tong et al. (2004) provided a genetetic interactions study using synthetic lethality. 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). Structural approaches employed modelling of interactions using structural templates derived from known protein complexes (Aloy et al., 2002, 2004).

In this article, we follow Rhodes et al., 2005 and Pospisil et al., 2006 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.

2 RESULTS

2.1 Approach

In this study, we design a computational approach to automatically reconstruct pathway maps and interaction networks of proteins. Applied to genes involved in pancreas cancer, we 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 1.

Gene expression data (1). Our data set [Fig. 1, (1)] was obtained by integrating our various analyses of the gene expression profiles of PDAC from Affymetrix GeneChip experiments such as microdissection, systematic isolation of genes (Grutzmann et al., 2003a, b, 2004b), and the meta-analysis of PDAC gene expression profiles from publicly available data (Grutzmann et al., 2005). The data set pooled from these studies contains 1612 genes differentially expressed in pancreatic ductal adenocarcinoma (PDAC).

From expression to pathways (2,3). 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 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.

Known interactome by localization (6). We obtain all experimentally known interactions within our data set...
Structural templates predict novel interactions and targets

Fig. 2. Overview of the 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.

from the literature (4) by help of the NetPro database (http://www.molecularconnections.com/protein_interactions.html). For every protein, we then retrieve localization information from the Gene Ontology cellular component annotation (5). From this, we construct an integrative map of known pancreas cancer relevant protein interactions (6) (data not shown).

Structure-based interaction predictions (7–9). 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 uncharacterized cancer related genes. Our interaction prediction approach is based on SCOPPI, the Structural Classification of Protein-Protein Interfaces (Winter et al., 2006). The idea of predicting new interactions from these known ones is sketched in Figure 1 on the right (see ‘Methods’ for details). The resulting set of initial interaction predictions [Fig. 1, (8)] yields ~1,000 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 40 confident, novel interactions. Table 1 contains the subset of 29 interactions where only interactions between a pair of up-up or up-down regulated genes are shown in addition to two literature confirmed interactions which are down-down regulated. A table with all 40 interaction is provided as supplementary material.

A pancreas cancer map (10). By linking 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 PDAC data that are involved in the 40 novel predicted interactions.

Validation of candidates (11,12). 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. Molecular Dynamics simulations confirmed that the predicted TMPRSS4-TFPI2 interaction remains stable. We further performed docking experiments which indicate that BVDU (RP101) is able to bind to the active site of TMPRSS4.

3 DISCUSSION

3.1 Cancer genes

Around 1,500 of the genes in our data set are validated by checking them against previously reported differentially expressed cancer genes (Higgins et al., 2007), among them the K-ras oncogene whose mutation has been identified in 90% of pancreas cancers, the insulin-like growth factor (IGFBP4/5), STAT1 from the signal transducers and activators of transcription family.

SMADS are proteins of the TGFBs signalling pathway. Sova et al. (2006) identified TFPI2 as a biomarker that is hypermethylated and repressed in cervical cancer. TMPRSS4 has been also identified as a biomarker for thyroid cancer. 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.

Downregulation or loss of SMAD4 was shown to be important for pancreatic carcinogenesis. 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 (Kato et al., 2005). BRCA1, whose mutation appears to confer increased susceptibility for PDAC (Hezel et al., 2006), 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.

Some of the above identified genes were investigated as therapeutic targets. Fleming et al. (2005) 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. (2005) identified RAB6KIFL as a candidate for development of drugs to treat PDACs. Knockdown of endogenous RAB6KIFL 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. From our data, we can predict a potential interaction of RAB6KIFL and RAB22A.
Table 1. 29 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 SCOPPI. 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 $\checkmark$. The $\times$ sign represents a negative literature confirmation (Loll et al., 2004).

<table>
<thead>
<tr>
<th>Protein 1</th>
<th>Description</th>
<th>up/down</th>
<th>Interface conserved</th>
<th>Complex PDB ID</th>
<th>Protein 2</th>
<th>Description</th>
<th>up/down</th>
<th>Interface conserved</th>
<th>Confirmed by literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRM</td>
<td>Myosin regulatory light chain 2, nonsarcomeric</td>
<td>up</td>
<td>63%</td>
<td>1b7t</td>
<td>MYH9</td>
<td>Cellular myosin heavy chain, type A</td>
<td>up</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>TFPI2</td>
<td>Tissue factor pathway inhibitor 2</td>
<td>down</td>
<td>62%</td>
<td>1uw</td>
<td>KLK10</td>
<td>Kallikrein 10</td>
<td>up</td>
<td>59%</td>
<td></td>
</tr>
<tr>
<td>TFPI2</td>
<td>Tissue factor pathway inhibitor 2</td>
<td>down</td>
<td>61%</td>
<td>1bc</td>
<td>TMPRSS4</td>
<td>Transmembrane protease, serine 4</td>
<td>up</td>
<td>63%</td>
<td></td>
</tr>
<tr>
<td>TMPRSS4</td>
<td>Transmembrane protease, serine 4</td>
<td>up</td>
<td>64%</td>
<td>1ezx</td>
<td>SERPIN12</td>
<td>Protease inhibitor 14</td>
<td>down</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>TFPI2</td>
<td>Tissue factor pathway inhibitor 2</td>
<td>down</td>
<td>61%</td>
<td>1b7t</td>
<td>MYH9</td>
<td>Cellular myosin heavy chain, type A</td>
<td>up</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>RHOA</td>
<td>Transforming protein RhoA</td>
<td>up</td>
<td>89%</td>
<td>1am4</td>
<td>DLC1</td>
<td>Deleted in liver cancer 1, isoform 1</td>
<td>down</td>
<td>58%</td>
<td></td>
</tr>
<tr>
<td>RHOA</td>
<td>Transforming protein RhoA</td>
<td>up</td>
<td>80%</td>
<td>1kgf</td>
<td>Pleckstrin-2</td>
<td>down</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FYN</td>
<td>FYN Tyrosine Kinase protooncogene</td>
<td>up</td>
<td>77%</td>
<td>2src</td>
<td>EPS8L1</td>
<td>Epidermal growth factor receptor kinase sub-strate 8-like protein 1</td>
<td>down</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>FYN</td>
<td>FYN Tyrosine Kinase protooncogene</td>
<td>up</td>
<td>77%</td>
<td>2src</td>
<td>BIN1</td>
<td>Myc box-dependent-interacting protein 1</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>Complement component C2</td>
<td>up</td>
<td>64%</td>
<td>1ezx</td>
<td>SERPIN12</td>
<td>Protease inhibitor 14</td>
<td>down</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>Complement component C2</td>
<td>up</td>
<td>50%</td>
<td>1sgf</td>
<td>NTF5</td>
<td>Neurotrophin-5</td>
<td>down</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>KLK10</td>
<td>Kallikrein 10</td>
<td>up</td>
<td>57%</td>
<td>1ezx</td>
<td>SERPIN12</td>
<td>Protease inhibitor 14</td>
<td>down</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>RAB25</td>
<td>Ras-related protein Rab-25</td>
<td>up</td>
<td>56%</td>
<td>1jeq</td>
<td>ADCY9</td>
<td>Adenylate cyclase type 9</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>RAB25</td>
<td>Ras-related protein Rab-25</td>
<td>up</td>
<td>56%</td>
<td>1jeq</td>
<td>ADCY3</td>
<td>Adenylate cyclase type 3</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>RRAS</td>
<td>Ras-related protein R-Ras</td>
<td>up</td>
<td>100%</td>
<td>1wq1</td>
<td>RASAL2</td>
<td>RAS protein activator-like 2</td>
<td>up</td>
<td>59%</td>
<td></td>
</tr>
<tr>
<td>MYL9</td>
<td>Myosin, light polypeptide 9, regulatory</td>
<td>up</td>
<td>61%</td>
<td>1b7t</td>
<td>MYH9</td>
<td>Cellular myosin heavy chain, type A</td>
<td>up</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>KRAS2</td>
<td>GTPase KRas</td>
<td>up</td>
<td>100%</td>
<td>1wq1</td>
<td>RASAL2</td>
<td>RAS protein activator-like 2</td>
<td>up</td>
<td>59%</td>
<td></td>
</tr>
<tr>
<td>RGS2</td>
<td>Regulator of G-protein signalling 2, 24kDa</td>
<td>down</td>
<td>53%</td>
<td>1fqi</td>
<td>RAB22A</td>
<td>Ras-related protein Rab-22A</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>RGS5</td>
<td>Regulator of G-protein signalling 5</td>
<td>up</td>
<td>53%</td>
<td>1fqi</td>
<td>RAB22A</td>
<td>Ras-related protein Rab-22A</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>RGS16</td>
<td>Regulator of G-protein signalling 16</td>
<td>up</td>
<td>53%</td>
<td>1fqi</td>
<td>RAB22A</td>
<td>Ras-related protein Rab-22A</td>
<td>up</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>CDC2L1</td>
<td>Cell division cycle 2-like 2</td>
<td>up</td>
<td>67%</td>
<td>1qf1</td>
<td>CDKN3</td>
<td>Cyclin-dependent kinase inhibitor 3</td>
<td>up</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>RAB22A</td>
<td>Ras-related protein Rab-22A</td>
<td>up</td>
<td>56%</td>
<td>1jq2</td>
<td>KIF20A</td>
<td>Kinesin family member 20A</td>
<td>up</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>CDC2</td>
<td>Cell division control protein 2 homolog</td>
<td>up</td>
<td>83%</td>
<td>1fqi</td>
<td>CDKN3</td>
<td>Cyclin-dependent kinase inhibitor 3</td>
<td>up</td>
<td>100%</td>
<td>✓</td>
</tr>
<tr>
<td>CDK7</td>
<td>Cyclin-dependent kinase 7</td>
<td>up</td>
<td>58%</td>
<td>1fqi</td>
<td>CDKN3</td>
<td>Cyclin-dependent kinase inhibitor 3</td>
<td>up</td>
<td>100%</td>
<td>✓</td>
</tr>
<tr>
<td>ARHGDIA</td>
<td>Rho GDP dissociation inhibitor(GDI)alpha</td>
<td>up</td>
<td>100%</td>
<td>1ce0</td>
<td>RHOA</td>
<td>Transforming protein RhoA</td>
<td>up</td>
<td>100%</td>
<td>✓</td>
</tr>
<tr>
<td>EPS15L1</td>
<td>Epidermal growth factor receptor pathway substrate 15-like 1</td>
<td>down</td>
<td>50%</td>
<td>1dkf</td>
<td>MYH9</td>
<td>Cellular myosin heavy chain, type A</td>
<td>up</td>
<td>59%</td>
<td></td>
</tr>
<tr>
<td>TAPBP</td>
<td>TAP binding protein</td>
<td>down</td>
<td>67%</td>
<td>2ig2</td>
<td>CD58</td>
<td>Lymphocyte function-associated antigen 3</td>
<td>up</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>TFPI2</td>
<td>Tissue factor pathway inhibitor 2</td>
<td>down</td>
<td>62%</td>
<td>1taw</td>
<td>F11</td>
<td>Coagulation factor XI</td>
<td>down</td>
<td>66%</td>
<td>✓</td>
</tr>
<tr>
<td>TFPI2</td>
<td>Tissue factor pathway inhibitor 2</td>
<td>down</td>
<td>62%</td>
<td>1taw</td>
<td>KLKB1</td>
<td>Kallikrein B, plasma (Fletcher factor) 1</td>
<td>down</td>
<td>66%</td>
<td>✓</td>
</tr>
</tbody>
</table>
### 3.2 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. (2003) and Kayed et al. (2006) suggest 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., 2006). Xu and Attisano (2000) presented 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 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 synthesize, package and secrete a variety of digestive enzymes. This process is regulated by neurotransmitters and hormones, both of which utilize 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 that 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 colonize new terrain in the body.

### 3.3 Hallmark interactions of pancreatic cancer

Combining pathways, known interactions and predicted interactions, we obtain the hallmarks of pancreatic cancer map (Fig. 3). Our data confirm 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 1 lists 29 structure-based interactions 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 pancreas cancer is discussed as follows.

**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. In pancreas cancer cells, 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 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 Fig. 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 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 minimization of the complex, the pocket around the active site slightly opens (Fig. 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 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 has been shown that TFPI2 inhibits plasmin, trypsin, chymotrypsin, cathepsin G and plasma kallikrein but not urokinase-type plasminogen activator, tissue plasmin and thrombin (Konduri et al., 2001). It plays a major role in negative regulation of the coagulation cascades (upper right in Fig. 2) and its downregulation is associated with malignant pancreas tumours. On the other hand, TMPRSS4 is known to be upregulated in pancreas 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.

**A proposed mechanism of action for BDVU.** For pancreas cancer, one of the standard drug treatments is gemcitabine-based chemotherapy. Recently, these standard chemotherapies were found to give better results when combined with specific substances sensitizing the tumour towards chemotherapy. The effect of BDVU ((E)-5-(2-bromovinyl)-2'-deoxyuridine...
or RP101), which supports apoptosis and prevents the acquisition of chemoresistance, was demonstrated in vitro and in patients with pancreas cancer (Fahrig et al., 2006). BVDU co-treatment significantly enhanced survival and time to progression. 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 speculate that BVDU might act as an inhibitor of TMPRSS4, replacing the function of downregulated TFPI2. To confirm this hypothesis, we used PatchDock (Schneidman-Duhovny et al., 2003) to dock BVDU to TMPRSS4. The result is shown in Figure 4f. BVDU clearly blocks the pocket with the active site. We can only speculate about the affinity of BVDU towards TMPRSS4, but we believe that it at least acts as competitive inhibitor for the natural substrate of TMPRSS4. This suggests a potential novel role of BVDU as a TMPRSS4 inhibitor.

Validation of predictions. We used homology modelling and docking to further test our results. For the homology modelling we use the MODELLER software for homology or comparative modelling of protein three-dimensional structures (Mart-Renom et al., 2000). The Modeller results for the TMPRSS4–TFPI2 interaction strongly supports our prediction. As input we provide an alignment of the TMPRSS4 sequence to be modelled with known related structures and the output is the modelled structure of the input sequence.

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.

Fig. 3. A comprehensive map of pancreas cancer relevant interactions and pathways. The underlying picture was taken from (Hanahan and Weinber, 2000) and updated with our findings. It depicts integrated circuit of the cell progress annotated with the PDAC genes that are involved in the 40 novel predicted interactions. Proteins are shown according to their cellular localization and their associated pathways. Genes coloured green are downregulated while genes coloured red are upregulated. The boxes around some of the genes indicates that those genes were present on the original map and the colours of the boxes indicate their expression level. Lines linking genes represent interactions among the genes. Interactions confirmed by literature shown in red, and predicted interactions shown in blue.
Limitations. For the interactions predicted from known complex structures (Table 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 METHODS

Data set. The gene expression data used in our work originates from the four microarray studies mentioned in the results section (Section 2.1). 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. Our data set consists of 1612 genes, out of which 944 were found to be up-regulated and 668 downregulated in PDAC.

Public resources. We want to briefly summarise the data resources used in our study. The KEGG Pathways database (Kanehisa et al., 2005) 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. NetPro is the proprietary protein interaction database covering more than 100,000 expert curated and annotated protein–protein interactions. All the interactions were obtained from peer reviewed published scientific articles and have gone through expert cross-checking quality checks. The Protein Data Bank, PDB (Berman et al., 2000) is a repository for three-dimensional structures. As of January 2007, it contains some 39,000 protein structures, most of which have been obtained by X-ray crystallography. Around half of the PDB entries are multi-domain protein structures. The structural classification of proteins, SCOP (Murzin et al., 1995) provides a hierarchical classification of protein structures at domain level. The hierarchy contains four levels: class, fold, superfamily and 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. The Genomic Threading Database, GTD (McGuflin et al., 2004) assigns structural folds to proteins of unknown structure. Structural annotations are carried out using a modified version of GenTHREADER (Jones, 1999). GTD is more sensitive than sequence alignment, and can assign folds correctly even with low sequence similarity. The Structural Classification of Protein-Protein Interfaces SCOPPI (Winter et al., 2006) is a database containing all domain-domain interactions and their interfaces of multi-domain proteins from the PDB which follows the rule: two domains are considered as interacting if there are at least 5 residue pairs within 5 Å.

Structure-based prediction of protein interactions. We implemented a methodology that utilises structural data from SCOPPI to predict potential interaction among the PDAC data set deregulated genes. The resulting potential interactions are further investigated by considering amino acid sequence conservation of ≥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 1: (i) Structure assignment and Family classification. Most of the data set genes are of unknown structure. First, we use the Genomic Threading Database (GTD) as fold recognition method to assign SCOP domain structures to all proteins in our data set. Only assignments with certain and high confidence by GTD are considered. This results in 656 remaining genes. (ii) Interaction prediction. For the assigned SCOP domains, we use SCOPPI to identify interacting domain pairs. In this step, we consider two proteins 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. Figure 1 sketches the structure assignment and interaction prediction step of the method. This initial interaction prediction is further refined. (iii) Interface conservation evaluation. It has been shown that protein interface residues are usually more conserved than the rest of the exposed surface (Elock 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 Å 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 or equal to 50% of residues at interface against counterpart of the known template structure, we assume that they share the same interaction partner. Applying this criterion to the whole PDAC data, many interactions are filtered out, and 40 remain. (iv) Interaction confirmation. In order to evaluate our method, we compared our finally predicted interactions against those confirmed by experimental interaction databases. For this purpose, NetPro, BIND, and HPRD (Peri et al., 2003) are used.

Modelling and Docking procedures. For the homology modelling we used MODELLER version 8v0 (Mart-Renom et al., 2000). BDOCK (Huang and Schroeder, 2005) was used for docking. We applied conjugate gradient energy minimization using NAMD (Phillips et al., 2005) with the CHARMM22 force field. For the simulation on the TMPRSS4–TFPI2 complex, we observed a stabilization of the complex after 10,000 steps. The structure of TMPRSS4–TFPI2 complex is provided as Supplementary Material.

Protein structures. The following structures were used from the Protein Data Bank: Complex of trypsin interacting with ankylosis 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 BDVU structure was taken from Crystal Structure of Thymidine Kinase from Herpes Simplex Virus Type I (PDB ID 1k8).

5 SUMMARY AND CONCLUSION

In this study, we propose an integrative approach 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.

Second, we propose a method that predicts interactions among a given set of genes. 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 pancreas cancer genes. As a result, we predict 40 novel interactions that are specific for the underlying disease. We map these interactions onto a well-known picture of cancer hallmark 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 pancreas cancer.

Finally, we examine one example in detail: 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. BVDU is known to enhance survival time in patients with

5 SUMMARY AND CONCLUSION

In this study, we propose an integrative approach 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.

Second, we propose a method that predicts interactions among a given set of genes. 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 pancreas cancer genes. As a result, we predict 40 novel interactions that are specific for the underlying disease. We map these interactions onto a well-known picture of cancer hallmark 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 pancreas cancer.

Finally, we examine one example in detail: 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. BVDU is known to enhance survival time in patients with
pancreatic cancer. We hypothesise that BVDU can bind to the active site of TMPRSS4 and thus acts as its inhibitor.

ACKNOWLEDGEMENTS

We thank Andreas Henschel for his help with modelling and for fruitful discussions. We further would like to thank Bingding Huang for his help on docking, and Anne Tuukkanen for energy minimization experiments.

Parts of Figure 3 reprinted from Cell, Vol. 7, D. Hanahan and R. A. Weinberg, The hallmarks of cancer, 57–70, Copyright (2000), with permission from Elsevier.

Conflict of Interest: none declared.

REFERENCES