Chapter 9
Interactomics and Cancer

Gautam Chaurasia and Matthias E. Futschik

Abstract Cancer is a complex disease with a myriad of genes and molecular processes involved. To unravel its underlying mechanisms, the main approach to date has been the study of individual genes and their association with carcinogenesis. As a recently emerging new paradigm, systems biology has complemented this time-honoured concept by promoting a holistic view of cancer as a network-associated disease. This new strategy is reflected par excellence by the construction of genome- and proteome-wide interaction networks and their utilization. We give here an overview of the current status of the human interactome and report first successes in its application in cancer research. In particular, interactomics-based analyses have been successfully undertaken for the characterization and de novo prediction of cancer-associated genes and processes. Although considerable challenges are still to overcome, interactomics promises to become a cornerstone in the systems biology of cancer.

9.1 Introduction

Cancer is not a single uniform disease, but displays a striking heterogeneity in its cause, progression and prognosis. In fact, more than 100 distinct types of cancer have been identified in a variety of tissues over the last decades (Hanahan and Weinberg 2000). The recent progress in molecular profiling of cancer is likely to contribute to an even larger number of biologically and clinically distinct tumor sub-types (Alizadeh et al. 2000). Such observed heterogeneity is not only of interest for cancer researchers, but has also direct consequences in the clinical prognosis and medical treatment of cancer patients.

G. Chaurasia
Charité, Humboldt-University, Berlin, Germany

M.E. Futschik (✉)
Centre for Molecular and Structural Biomedicine, University of Algarve, Faro, Portugal
e-mail: mfutschik@ualg.pt

W.C.S. Cho (ed.), *An Omics Perspective on Cancer Research*,
DOI 10.1007/978-90-481-2675-0_9,

Where does the observed heterogeneity originate from? Intensive research has discovered a large number of genes involved in the development of cancer. Especially, the study of genetic mutations identified many cancer-associated genes and has led to the view of cancer as a primarily genetic disease. A recent census of human cancer genes showed that somatic and germline mutations in almost 400 genes have repeatedly been reported to contribute to oncogenesis (Futreal et al. 2004). Additionally, numerous epigenetic and transcriptional changes have been associated with cancer (see also chapters 4 and 5).

How can we cope with this complex and heterogeneous disease in which so many genes and processes are involved? For a long time, the main approach to unravel oncogenesis has been to identify single cancer-associated genes and to characterize them one at a time. Undoubtedly, this paradigm in cancer research has supplied us with an impressive catalogue of pathogenic changes on molecular level. Despite considerable success, however, it has not yet delivered the much anticipated "magic bullets" against this disease.

Recently, a new discipline has emerged with the advent of large-scale biological data sets: *Systems biology*. It can be viewed as a complementary – but not opposing – approach to the classical reductionistic strategy for the study of the biological processes. In contrast to reductionistic approaches based on the dissection of processes into their most elementary levels, systems biology is more holistically orientated. The guiding principle of systems biology is that the total system can be more than the sum of its parts and can acquire properties that are not implied in the single components.

Following this principle, we seek to study a biological system as a whole. The aim is to determine the rules governing its behaviour and eventually to generate qualitative and quantitative predictions concerning its response to perturbations and modifications. To achieve this, two requirements have to be fulfilled: (1) a sufficient amount of data and information describing the system has to be available, and (2) a computational model of the system has to be designed. Whereas the first requirement is increasingly met with the development of new high-throughput techniques, the second one still demands considerable efforts. For instance, when we aim to represent the whole system, we need to choose an adequate level of resolution. Finding this level is challenging, since there is usually a trade-off between computational feasibility and detailed representation of the molecular systems due to their mere size and complexity. The inclusion of too many components can lead to ill-determined models of the system with many parameters unknown, whereas a too severe restriction can results in an incomplete model with a lack of coherence. In fact, the choice of a suitable model depends not only on the research objective, but also, more practically, on the quality and quantity of data and information present.

In response to this difficulty, various methodologies for different levels of resolution have been brought forward in systems biology to date. A nowadays very popular approach is based on the representation of biological systems as mathematical graphs and has laid the ground for the blooming field called *network biology*. In the context of molecular systems, for instance, the molecules are typically

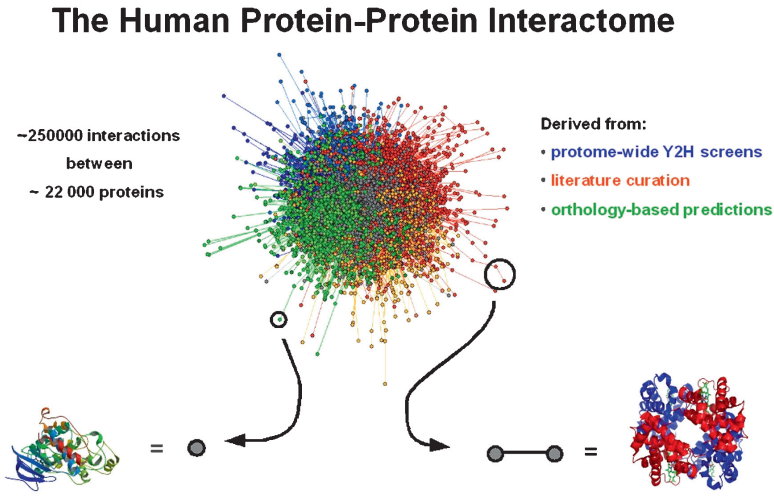

Fig. 9.1 Graphical representation of the current human protein–protein interactome as stored in the UniHI database (http://www.unihi.org)**.** Altogether, it comprises over a quarter of a million of interactions derived from experimental resources and by computational prediction. Nodes and edges in the displayed graph represent proteins and their interactions. The different colours indicate the source of interactions: blue - Y2H screens, red - literature curation, green - orthology-based prediction, and grey - multiple evidence. Notably, distinct regions of the interactome are covered by different methods indicating the potential benefits of integration. The figure also illustrates the grade of simplification achieved by the graph-theoretical approach. The highlighted nodes symbolizing the shown proteins (*left*: mitogen activated protein kinase; *right*: haemoglobin complex consisting of alpha and beta chains) are depicted for illustration only; they do not represent the actual location of these proteins in the interactome. Displayed protein structures were taken from the Protein Data Bank

represented as nodes and their interactions as edges (Fig. 9.1). Although this type of representation is clearly a stark simplification of the underlying physical system, a major advantage of this approach is that the analysis of large networks becomes feasible. Also, the underlying graph-theory has been well developed and offers researchers a variety of tools. In fact, with its beginning dating back to Leonard Euler in 1736, graph theory has made profound impact in social, physical and computer sciences (Euler 1736). The application of graph-theory to biology seems to be well suited where large networks are involved in the process of interest. Thus, it is not surprising that the concepts of network biology have been especially applied to elucidate the complex processes during oncogenesis and to consolidate the hitherto divergent observations. A short introduction to graph theory and its basic concepts is presented in Box 1.

The reminder of this chapter is following: We first present an overview of current strategies to chart, to store and to analyse interaction maps. We focus here on protein–protein interaction data as many concepts of network biology have originally been demonstrated using protein interactions. Notably, we describe the generation of protein interaction maps in some details, as it can have a considerable

Box 1 Introduction to Graph Theory and Its Application to Network Biology

Graph-Theoretical Description of Molecular Networks

One of the most basic descriptions of molecular systems is given by their representation as mathematical graphs. For protein interaction networks, for instance, proteins are commonly represented as nodes and their physical interactions as undirected edges. For transcriptional regulatory networks, nodes symbolize both transcription factors and their target genes and are connected by direct edges. The resulting graphs can be analyzed using various graph-theoretical measures:

A fundamental characteristic of a node in a mathematical graph is its degree, i.e. the number of edges to other nodes. The *degree distribution P(k)* for a network can be defined as fraction of proteins with *k* interactions in the total network. It is an important feature of network to distinguish different network classes. Of special importance here is the power-law distribution ($P(k) \sim k^{-\gamma}$) which is characteristic for the class of scale-free networks. It has been shown that such network architecture is more robust against random failure of single components. A consequence of the scale-free topology is the emergence of so-called network *hubs*, i.e. highly connected nodes. Hubs are of particular importance for the network integrity and were associated with essential proteins (Jeong et al. 2001). Finally, the *shortest path* length between two nodes is defined as the minimum number of edges included in the (directed) path between the two nodes.

influence on the final maps. In fact, it is of critical importance for researchers to have a basic understanding how interaction maps were derived to avoid pitfalls in their usage. Subsequently, several studies and methodologies utilizing protein interaction networks to study cancer are reviewed. For sake of completeness, some references to the application of transcriptional networks to cancer research are given. Finally, we discuss future challenges and directions in the generation of human protein interaction maps and their applications.

9.2 The Human Protein–Protein Interactome: Generation and Analysis

In the last few years, we have witnessed the rapid increase in the large-scale protein–protein interaction maps for various model organisms. This striking rise is mainly due to advances in the high-throughput experimental techniques such as Yeast-two-Hybrid (Y2H), the coordinated efforts to systematically chart interactions by human experts as well as the progress in computational text-mining and prediction.

As all these methods can lead to considerably divergent protein interaction maps (von Mering et al. 2002; Futschik et al. 2007a, b), it is important to have a basic understanding of the applied methodologies. In the following sections, we therefore introduce several current methods, discuss their pros and cons and outline their application to the human interactome.

9.2.1 Yeast-Two Hybrid System

The Y2H method is based on a screening approach using a set of modified proteins. The experimental basis of Y2H is the reconstitution of a multi-domain transcription factor (such as GAL4). Specifically, a protein-encoding cDNA of interest is cloned into a bait vector, and fused with the DNA binding domain of the multi-domain transcription factor. A second cDNA encoding a potentially interacting protein is cloned into a prey vector and fused to the transcription factor's activation domain. Subsequently, the two yeast strains carrying the bait and prey hybrid proteins in plasmids are mated, resulting in yeast carrying both plasmids. If the bait and prey proteins interact, a functional transcription factor is reconstituted leading to the transcription of a reporter gene such as lacZ encoding for β-galactosidase. In the high-throughput mode, whole libraries of bait and prey vectors can be screened for interactions. Thus, the main advantage of this approach is that it provides a platform for the rapid generation of large-scale protein–protein interaction networks and it does not need to be biased towards known interactions. However, the false positive rate for Y2H screens can be considerable and can even exceed the estimated true positive rates (Hart et al. 2006).

Recently, the Y2H system was applied in two large-scale studies to screen human proteins identifying in total over ~5,500 new protein interactions, of which a selected sub-set was experimentally validated (Rual et al. 2005; Stelzl et al. 2005). Notably, the overlap between the two studies is small: Only 17% of interactions between common proteins were detected by both groups.

9.2.2 Literature Curation and Text-Mining

Besides high-throughput experimental approaches, the numerous small-scale experiments described in the literature can be exploited to create large-scale protein interaction maps. Tapping into the wealth of published experiments, information on protein interactions is systematically extracted from the literature either by human experts or text-mining algorithms. The advantages of such procedures are that it is not biased *a priori* towards a particular experimental technique and that the charted interactions are determined under a broad range of conditions and protocols. Characteristic disadvantages are the inherent difficulty to estimate the false positive rate and the bias towards highly studied proteins. Numerous research groups have

followed this strategy to create large-scale human protein interaction maps (Bader et al. 2001; Salwinski et al. 2004; Pagel et al. 2005; Ramani et al. 2005; Mishra et al. 2006; Kerrien et al. 2007; Breitkreutz et al. 2008).

9.2.3 Computational Prediction of Human Protein Interactions

Alternative to the large-scale experimental and literature-curation, *in silico* prediction has been used to build large-scale protein–protein interaction maps (Lehner and Fraser 2004; Brown and Jurisica 2005; Persico et al. 2005). This strategy is based on the assumption that interactions are evolutionarily conserved for orthologous proteins and thus interactions detected between proteins in lower organisms can be extrapolated to their human orthologs. A main advantage of this method is that it is entirely computational and thus enables rapid and cost-effective construction of human protein–protein interaction maps. Disadvantages are that it is purely predictive in nature and false positives can arise through erroneous mapping to human orthologs or that interactions are simply lost during evolution.

9.2.4 Databases for Human Protein Interactions

Several human protein interaction databases have been established to help researchers find and analyze interaction partners of proteins of interest. These databases can generally be divided into two different categories: The first one is based on the manual-curation of published literature and includes the Human Protein Reference Database (HPRD), the Biological General Repository for Interaction Datasets (BioGRID), IntAct, the Database of Interacting Proteins (DIP), the Biomolecular Interaction Network Database (BIND) and the MIPS Mammalian Protein–Protein Interaction Database (MPPI) (Bader et al. 2001; Salwinski et al. 2004; Pagel et al. 2005; Mishra et al. 2006; Kerrien et al. 2007; Breitkreutz et al. 2008). The other category of databases also includes computationally predicted interactions; examples of such databases are the Online Predicted Human Interaction Database (OPHID) and HomoMINT (Brown and Jurisica 2005; Persico et al. 2005). Currently, HPRD is one of the major sources for human interaction data and – as the name implies – dedicated to human proteins. Besides interactions, it also provides information on domain architecture, post-translational modifications, disease association and biological pathways. Other databases e.g. BioGRID, IntAct, DIP and BIND are the repositories for a more diverse set of organisms and provide access to interaction data for other model organisms such as yeast, worm and fly.

Although these databases chart thousand of interactions from human proteins, their coverage in terms of the whole human interactome remains rudimentary. Comparative analysis revealed a very limited overlap between them (Chaurasia et al. 2006; Futschik et al. 2007a; Ramírez et al. 2007) (Fig. 9.1). Naturally, the

question arises why these maps have such small degree of overlap. One reason is likely that current maps are highly unsaturated. Given an estimated total size of human interactome of ~650,000 interactions, even HPRD – as one of the largest sources – covers less than 10% of the total interactome (Stumpf et al. 2008). Additionally, current maps display a strong detection bias, i.e. they are enriched in characteristic types of proteins while depleted of other types (Futschik et al. 2007a). For example, literature-based maps show a significant enrichment in signalling proteins which is probably due to their popularity as biomedical research topic. Since currently available maps are incomplete and might contain complementary information, we and others reasoned that their integration can be beneficial. Therefore, several research groups have started to integrate the diverse protein interaction datasets available (Prieto and De Las Rivas 2006; Chaurasia et al. 2007). For instance, the Unified Human Interactome database (UniHI) including human interaction data from 14 different sources stores over ~250,000 interactions between ~22,000 proteins and thus constitutes one of the most comprehensive collections of human protein interactions at present (Chaurasia et al. 2009). Such centralized repositories liberate researchers from laborious and time-consuming integration of the diverse interaction data sets. An overview of several current resources for human protein interactions is provided in Table 9.1. A more complete list of protein interaction databases is compiled by the Pathguide project (Bader et al. 2006).

9.3 Application of Interactomics to Cancer Research

9.3.1 Network-Based Characterization of Cancer Genes

One of the first questions addressed by network-based approaches in cancer research is also one of the most intriguing: What makes a gene to a cancer gene? Although such naïve question may be somewhat puzzling at first, it makes naturally sense in network biology to ask whether cancer-associated genes have characteristic properties within interaction networks. To address this question, graph-based methods can be applied to study network properties of cancer genes. An important concept here is *centrality* which evaluates the location within a network. Centrality of a node can be defined simply by its degree, i.e. the number of interactions or, more elaborately, by the number of shortest paths passing through this node.

Several research groups have applied such concepts to reveal the graph-theoretical properties and the role of cancer genes in human protein interaction networks (Wachi et al. 2005; Jonsson and Bates 2006; Hernández et al. 2007; Platzer et al. 2007). For the analysis, the set of cancer-associated genes has first to be determined, for which commonly databases or microarray studies are used. As a second step, a disease network is created by integrating the cancer genes products (i.e. proteins encoded by cancer-associated genes) with available large-scale protein

Table 9.1 Resources for human protein–protein interactions described in the chapter. The size, the construction approach and additional information are given. For the calculation of the number of proteins and interactions in each dataset, proteins were mapped to their corresponding Entrez Gene identifiers

Resource	Proteins	Interactions	Method	References	Resource location
MDC-Y2H	1,703	3,186	Y2H SCREEN	Stelzl et al. (2005)	www.mdc-berlin.de/neuroprot
CCSB-Y2H	1,549	2,754	Y2H SCREEN	Rual et al. (2005)	www.vidal.dfci.harvard.edu (flat file only)
HPRD-BIN	8,788	38,800	LITERATURE	Peri et al. (2003)	www.hprd.org
DIP	1,085	1,397	LITERATURE	Salwinski et al. (2004)	www.dip.doe-mbi.ucla.edu
BIOGRID	7,953	24,624	LITERATURE	Breitkreutz et al. (2008)	www.thebiogrid.org
INTACT	7,273	19,404	LITERATURE	Hermjakob et al. (2004)	www.ebi.ac.uk/intact
BIND	5,286	7,394	LITERATURE	Bader et al. (2001)	www.bind.ca
COCIT	3,737	6,580	TEXT MINING	Ramani et al. (2005)	www.Bioinformatics.icmb.utexas.edu/idserve
REACTOME	1,554	37,332	LITERATURE	Joshi-Tope et al. (2005)	www.reactome.org
ORTHO	6,225	71,466	ORTHOLOGY	Lehner and Fraser, (2004)	www.sanger.ac.uk/PostGenomics/signaltransduction/interactionmap
HOMOMINT	4,127	10,174	ORTHOLOGY	Persico et al. (2005)	www.mint.bio.uniroma2.it
OPHID	4,785	24,991	ORTHOLOGY	Brown and Jurisica (2005)	www.ophid.utoronto.ca
UniHI	22,307	200,473	INTEGRATION	Chaurasia et al. (2009)	www.unihi.org

interaction networks. Finally, the topological properties (e.g. degree distribution, centrality) of cancer genes within this network are computed and compared to those of genes that have not been associated with cancer.

Wachi and co-worker applied the above outlined strategy to study the centrality of genes that are differentially expressed in cancer (Wachi et al. 2005). For their analysis, human interaction data was collected from OPHID. Microarray data were obtained from five patients with squamous cell lung cancer and compared to normal

samples of the same patients. Using a paired *t*-test, differentially regulated genes were determined and mapped onto the protein interaction network. The subsequent analysis revealed that up-regulated genes tend to be highly connected and more centrally located in the network compared to randomly selected genes. Down-regulated genes tended to be also more highly connected but not significantly. Furthermore, they did not show an increased centrality. Based on their findings, the authors suggested that a core set of central genes has to be activated during the course of carcinogenesis.

Similar results were reported in a separate topological analysis performed by Jonsson and Bates (2006). In contrast to Wachi et al., this analysis did not depend on microarray experiments to define cancer-associated genes. To avoid a bias towards a particular cancer type, they selected a general set of cancer genes that were previously identified in a literature-based census (Futreal et al. 2004). The human interaction network was constructed using an orthology-based approach. After mapping of cancer genes onto the human protein interaction network, the connectivity of each protein in the integrated network was computed. Results indicate that the cancer proteins show higher degrees than non-cancer proteins. Cancer proteins also tend to function as central hubs, reflecting their role as a key player in protein–protein interaction network. Clustering analysis additionally showed that cancer proteins, on average, are more frequently located in the interfaces between clusters indicating an enhanced role in the coordination of different cellular processes.

Following the same strategy as Wachi et al., Hernández and colleagues reported somewhat contrasting results for the topological properties and organization of cancer gene products in the human interactome network (Hernández et al. 2007). They started their analysis by creating an integrated set of human interactome originated from five manually-curated literature-based dataset. Microarray data sets for prostate, lung and colorectal cancers were utilized and differential expression was calculated. Topological analysis of the integrated network revealed that down-regulated genes consistently tend to be more centrally located. In contrast, the centrality of up-regulated genes was dependent on the chosen cancer type. They also found that topological properties of down-regulated cancer genes are correlated with common biological processes and pathways that lead to cancer. However, both types of genes appear to be important for the organization and integrity of network structure. In particular, the elimination of cancer-associated genes from the network results in a faster breakage of the original network in smaller networks than those observed for elimination of randomly chosen genes.

Finally, the most comprehensive graph-theoretical study for cancer to date was conducted by Platzer et al. (2007). Altogether, they analysed 29 genome-wide cancer expression data sets using 22 individual graph-theoretical measures. For each study, differential gene expression was determined and sub-graphs of differentially regulated genes were constructed based on interaction data from OPHID. Various properties of the sub-graphs such as size, modularity and density were subsequently examined. The main result was that genes showing differential expression in cancer

tend to interact and to form larger sub-networks than expected by chance. Strikingly, however, the prevalence of hub proteins was not increased in cancer-associated subgraphs. The authors speculated that extended graphs with low density indicate networks of high robustness against the failure of single genes. This is especially intriguing in the context of cancer, as such finding would demand for the simultaneous therapeutic targeting of multiple proteins.

In summary, the described network studies give a first overview about the structural role of cancer genes in protein interaction networks. Nevertheless, care has to be taken in interpretation as current interaction maps often show divergence in structure due to different methods used for their assembly (Futschik et al. 2007b).

9.3.2 Identification of New Cancer-Associated Genes and Processes Using Protein Interaction Networks

A second area in which protein interaction networks have been utilized in cancer research is the identification of new cancer-associated genes. The rationale behind these investigations is that interacting proteins are likely linked to the same or similar phenotype. A leading example is Fanconi anemia, a genetic disease, for which seven of the nine associated proteins form a physical complex involved in DNA repair. Although interaction data can provide a suitable first basis for *de novo* identification of disease-causing genes, additional information has commonly been utilized to improve specificity.

For many years, genetic linkage studies were the most potent approach to find new disease-causing genes. A major difficulty, however, is to pick the right gene within extended chromosomal regions that have been linked to a disease. Oti et al. showed that this task can be considerably facilitated using protein interaction data (Oti et al. 2006). For genetically homogenous diseases, they predicted new disease associations when genes fell within an identified susceptibility locus and have protein interactions with a gene known to cause this disease. This simple method of data integration led to a tenfold increased specificity compared to randomly selected candidate genes at the same locus. Notably, Oti et al. also deduced that protein interactions added as much information as localization to the prediction accuracy. In a similar study, Franke et al. extended the protein interaction network by including microarray and gene annotation to generate a functional interaction network (Franke et al. 2006). Also, new candidate genes were identified in the larger network neighbourhood of known disease genes, avoiding the restriction to direct interactors only.

One requirement of these studies is that we have to know already a set of genes associated with a certain disease. This set can be then used to “anchor” a disease in the human interactome. If however no such genes are known, this approach cannot be used. To overcome this limitation, Lage et al. catalogued human phenotypes in a computationally tractable manner (Lage et al. 2007). Their motivation was that similar diseases might share the same molecular basis. Having defined a

score for the similarity of phenotypes, information for a specific disease can then be deduced from similar diseases. Thus, candidate genes can be predicted even if no other gene associated with the specific disease is known yet. For prediction, Lage et al. integrated human protein interaction with linkage data in a similar manner as Oti et al. and Franke et al. Using an *in silico* pull-down approach and the similarity of phenotypes, they extracted known and new complexes and predicted several novel candidate disease genes involved in disorders such as cancer, Alzheimer's, diabetes and coronary heart diseases. Detailed analysis for epithelial ovarian cancer lead to the identification of a new candidate gene, Fanconi anemia group D2 protein (*FANCD2*) placed in a complex with breast cancer type 1 susceptibility protein (*BRCA1*) and breast cancer type 2 susceptibility protein (*BRCA2*). This protein has been associated with different types of cancer, but not with epithelial ovarian cancer so far.

A conceptually similar network-based modelling approach was applied by Pujana et al. to predict new candidate genes involved in breast cancer (Pujana et al. 2007). They assumed that genes, which are functionally related or showed conserved co-expression across species, might cause a similar phenotype. To test their hypothesis, they created a cancer-specific network with four known breast cancer-associated genes: *BRCA1*, *BRCA2*, *ATM*, and *CHEK2*. Neighbours of each reference gene set were further ranked using a scoring system based on co-expression, phenotypic similarity, and genetic or physical interactions among orthologs of the proteins in other species. They identified a new gene (*HMMR*) that was found to be associated with an increased risk of breast cancer.

In addition to prediction of novel cancer-associated genes, interaction networks were also employed to unravel cancer-related molecular processes. As one example, Chuang et al. applied a network-based classifier to identify sub-networks as markers for breast cancer prognosis (Chuang et al. 2007). To find the sub-networks, they mapped the gene expression profiles of metastatic and non-metastatic patients on a human protein–protein interaction network. Subsequently, they computed activity scores of all associated members to rank the sub-network as a whole. Their finding showed that high scoring sub-networks were enriched in many cancer-related biological processes such as apoptosis, proliferation, tissue remodelling, signalling and survival. Their analysis also indicated that identified modules were more reproducible than individual genes selected without network information, and that they achieve a higher accuracy in the classification of metastatic *versus* non-metastatic tumors. Another advantage of this approach is that it also captures those genes which may have not been detected based on gene expression data alone. Such non-differentially expressed genes could be an integral part of a complex and be required for connecting high scoring proteins in a sub-network. In fact, Chuang et al. found that a large number of the identified network structures contained at least one protein that was not significantly expressed in metastasis while most of them served as a bridge between high scoring proteins in a sub-network. This integration provides the opportunity to analyze the relationships between members of the complexes, and also increases the accuracy of the overall prediction.

9.3.3 Analysis of Transcriptional Regulatory Networks in Cancer Research

Besides physical protein–protein interactions, transcriptional regulations have been analyzed in network biology to shed light on oncogenesis. The main building blocks of the constructed transcriptional regulatory networks are transcription factors and their target genes. In contrast to the protein interaction networks, the resulting graphs are directed, i.e. include edges directed from transcription factors to their target genes. Since transcription factors can be themselves target genes of other transcription factors, this wiring scheme can lead to large connected networks. The ultimate goal is to build models that can "explain" observed expression patterns in terms of the underlying regulatory networks. Such models would go beyond the simple description of expression changes and could eventually provide us with a causative framework. This has become particularly interesting in the context of microarray technologies that have enabled a rapid genome-wide monitoring of expression.

In particular for yeast, this line of investigation has proven to be fruitful in revealing regulatory principles that are not detectable from the mere analysis of expression data (Janga et al. 2008). Early studies, for instance, could link changes in the structure of regulatory networks to the type of external stimuli and the corresponding transcriptional response (Luscombe et al. 2004). Such impressive interrogations were made possible by the systematic experimental mapping of yeast transcription factor binding sites using Chromatin-Immunoprecipitation on chip (ChIP-chip) experiments. Unfortunately, the systematic experimental charting of human transcription factor binding sites is still at a very early phase with experiments being limited to a small number of transcription factors and cell types. At present, many collections of transcription factor binding sites for humans thus rely considerably on *in silico* matching between promoter regions and position weighted matrices describing the consensus binding sites of transcription factors. Further difficulties in the construction of comprehensive regulatory networks are (1) a high number of false positive predictions of transcription factor binding sites based on simple sequence matching, (2) the choice of an adequate size of human promoter regions, (3) the combinatorial action of transcription factors within *cis*-regulatory modules and (4) the influence of the – generally unknown – chromatin structure on the accessibility of binding sites.

Despite these challenges, first efforts have been undertaken to construct genome-wide regulatory networks for cancer research. Notably, Kluger and colleagues examined the topological properties of regulatory networks to characterize gene deregulation during tumorigenesis (Tuck et al. 2006). For construction of a regulatory network, they utilized a collection of transcription factors stored in the Transcription factor (TRANSFAC) database. Potential target genes were determined by position weighted matrices. The basal connectivity network was then intersected with co-expression of genes from different cancer microarray studies to obtain condition-specific regulatory networks. In the subsequent analysis, network features such as degree distributions were used to differentiate between diseased and healthy patent samples. Although no significant improvement of classification accuracy was achieved compared to

conventional microarray analysis, the procedure offered some valuable insights in the potential causative mechanisms of gene deregulation. Most intriguingly, genes that discriminate best between disease conditions tend to be highly localized on the transcriptional network. It is important to note that the applied strategy implies that expression levels of transcription factors can be proxies for their activity states. However, this might neglect important post-translational modifications.

An impressive project, which can also serve as a prime example for integrative network biology, is the assembly and analysis of the B-cell interactome by Califano and co-workers. This model of the molecular network for B-cells not only includes transcriptional regulatory, but also protein–protein and modifying post-translational interactions derived from a variety of experimental and computational resources. In the study by Mani et al., a strategy was developed to scrutinize the B-cell interactome for dysregulated interactions in three distinct types of lymphoma (Mani et al. 2008). In contrast to conventional microarray analysis focusing on the differential regulation of genes, the loss or gain of correlation between interacting genes was analyzed. Remarkably, the examination of dysregulated interactions pointed more clearly to the set of known genetic lesions than simple differential gene expression did. Furthermore, potential downstream effectors could be identified which would have been missed using gene expression alone. Notably, these results probably would not have been derived without the construction of a cell type-specific network.

9.4 Summary and Outlook

Cancer shows a striking complexity in the cellular mechanisms involved and, despite all successes in cancer research, the untangling of these interwoven processes remains one of the most formidable tasks in molecular biology and medicine. For a long time, genes and their implications in cancer were studied one at a time. This time-honoured strategy has now been complemented with systems-wide studies of disease-associated mechanisms. A central position in the new paradigm has taken the uprising field of network biology. Applied to biomedicine, diseases represent particular states of the underlying molecular network; a perspective that was already brought forward several decades ago by S. Kauffman (Kauffman 1993). Following his influential ideas, cancer can be perceived as attractor states that might display remarkable robustness. Although based mostly on theoretical reasoning, we might argue to view cancer as a network-associated disease which requires complex intervention for its treatment (Kitano 2007).

A pivotal role in this new system biological strategy will be the study of protein interaction networks. Proteins and their aberrant interactions have long been known to be crucial in oncogenesis. With the construction of comprehensive interaction maps, we are now approaching a stage where the influence of dysfunctional proteins can systematically be dissected and potential interventions designed. Despite its early successes and rapidly growing popularity, the application of interactomics requires some caution.

Interaction maps of molecular processes are frequently highly rudimentary. This is also the case for human protein–protein maps in spite of their impressive size. At present, they are still scanty and are likely to include a considerable number of false positives. These shortcomings of current protein interaction networks – as well as of other types of molecular networks – underline the necessity of integrate complementary data and information. In fact, only by constructing multi-dimensional datasets, one can harvest the full potential of protein interaction maps. At present, this is mainly performed by simple mapping of expression changes onto generic interaction maps extracted from databases. Notably, such simple strategies account poorly for the complex spatial and temporal aspects of carcinogenesis. One step towards a more accurate representation can be the creation of tissue-specific networks. This might be especially relevant for cancer research where the examination of genes can lead to contradictory results depending on the used experimental model. For instance, *RAS*, a classical oncogene, has been shown to function in a tumor suppressing manner under certain conditions indicating the importance of the molecular context (Zhang et al. 2001). Also, the usefulness of streamlined interaction networks has already been demonstrated by the described study of the B-cell interactome. Future molecular maps reflecting this complexity will provide highly valuable tools for biomedical research. Indeed, the integration of independent information concerning expression and localization has already been used for the identification of dynamic as well as constitutive protein modules (Futschik et al. 2007c).

To conclude, early applications have indicated the large potential of network biology in cancer research. Progress in experimental techniques and computational methods will continue to improve the coverage and sensitivity of interaction networks. A focus of interactomics – especially in its application to cancer research – will be on the combination of different types of networks, such as protein-protein, transcriptional regulatory and metabolic networks, to enable the creation of detailed molecular models of oncogenesis. Furthermore, the integration of interactions networks with the rich datasets generated by ongoing cancer-related sequencing, microarray or imaging projects is likely to provide us with molecular maps of unprecedented detail for the human organism in health and disease. Thus, network biology promises to contribute substantially to a better understanding of the complexity of cancer and eventually to its cure.

Acknowledgements We would like to thank Paulo Martel and Nuno dos Santos for their important contributions to this chapter.

References

Alizadeh AA, Eisen MB, Davis RE et al (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511

Bader GD, Donaldson I, Wolting C et al (2001) BIND – The biomolecular interaction network database. Nucleic Acids Res 29:242–245

Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. Nucleic Acids Res Database 34:D504–506

Breitkreutz BJ, Stark C, Reguly T et al (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res Database 36:D637–640
Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21:2076–2082
Chaurasia G, Herzel H, Wanker EE et al (2006) Systematic functional assessment of human protein–protein interaction maps. Genome Inform 17:36–45
Chaurasia G, Iqbal Y, Hänig C et al (2007) UniHI: an entry gate to the human protein interactome. Nucleic Acids Res Database 35:D590–594
Chaurasia G, Malhotra S, Russ J et al (2009) UniHI 4: new tools for query, analysis and visualization of the human protein–protein interactome. Nucleic Acids Res Database 37:D657–D660
Chuang HY, Lee E, Liu YT et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
Euler L (1736) Solutio problematis ad geometriam situs pertinentis, Commentarii academiae scientiarum Petropolitanae 8:128–140
Franke L, van Bakel H, Fokkens L et al (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78:1011–1025
Futreal PA, Coin L, Marshall M et al (2004) A census of human cancer genes. Nat Rev Cancer 4:177–183
Futschik ME, Chaurasia G, Herzel H (2007a) Comparison of human protein–protein interaction maps. Bioinformatics 23:605–611
Futschik ME, Tschaut A, Chaurasia G et al (2007b) Graph-theoretical comparison reveals structural divergence of human protein interaction networks. Genome Inform 18:141–151
Futschik ME, Chaurasia G, Tschaut A et al (2007c) Functional and transcriptional coherency of modules in the human protein interaction network. J Integr Bioinform 4:76
Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100:57–70
Hart GT, Ramani A, Marcotte E (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7:120
Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington et al (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32: D452–D455
Hernández P, Huerta-Cepas J, Montaner D et al (2007) Evidence for systems-level molecular mechanisms of tumorigenesis. BMC Genomics 8:185
Janga SC, Collado-Vides J, Babu MM (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. Proc Natl Acad Sci USA 41:15761–15766
Jeong H, Mason SP, Barabasi AL et al (2001) Lethality and centrality in protein networks. Nature 411:41–42
Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22:2291–2297
Joshi-Tope G, Gillespie M, Vastrik I, et al (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res., 33:D428–D443
Kauffman SA (1993) Origins of order: self-organization and selection in evolution. Oxford University Press, New York
Kerrien S, Alam-Faruque Y, Aranda B et al (2007) IntAct – open source resource for molecular interaction data. Nucleic Acids Res Database 35:D561–565
Kitano H (2007) Towards a theory of biological robustness. Mol Syst Biol 3:137
Lage K, Karlberg EO, Størling ZM et al (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25:309–316
Lehner B, Fraser BA (2004) A first-draft human protein-interaction map. Genome Biol 5:R63
Luscombe NM, Babu MM, Yu H et al (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–312
Mani KM, Lefebvre C, Wang K et al (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 4:169
Mishra GR, Suresh M, Kumaran K et al (2006) Human protein reference database – 2006 update. Nucleic Acids Res Database 34:D411–414

Oti M, Snel B, Huynen MA et al (2006) Predicting disease genes using protein–protein interactions. J Med Genet 43:691–698

Pagel P, Kovac S, Oesterheld M et al (2005) The MIPS mammalian protein–protein interaction database. Bioinformatics 21:832–834

Peri S, Navarro JD, Amanchy R et al (2003) Development of human protein reference database as an initial platform for approaching 245 systems biology in humans. Genome Res., 13:2363–2371.

Persico M, Ceol A, Gavrila C et al (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6:S21

Platzer A, Perco P, Lukas A et al (2007) Characterization of protein-interaction networks in tumors. BMC Bioinformatics 8:224

Prieto C, De Las Rivas J (2006) APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res Web Server 34:W298–302

Pujana MA, Han JJ, Starita LM et al (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 39:1338–1349

Ramani A, Bunescu R, Mooney RJ et al (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol 6:R40

Ramírez F, Schlicker A, Assenov Y et al (2007) Computational analysis of human protein interaction networks. Proteomics 7:2541–2552

Rual JF, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437:1173–1178

Salwinski L, Miller CS, Smith AJ et al (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res Database 32:D449–D451

Stelzl U, Worm U, Lalowski M et al (2005) A human protein–protein interaction network: a resource for annotating the proteome. Cell 122:957–968

Stumpf M, Thorne T, de Silva E et al (2008) Estimating the size of the human interactome. Proc Natl Acad Sci USA 105:6959–6964

Tuck DP, Kluger HM, Kluger Y (2006) Characterizing disease states from topological properties of transcriptional regulatory networks. BMC Bioinformatics 7:236

von Mering C, Krause R, Snel B et al (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417:399–403

Wachi S, Yoneda K, Wu R (2005) Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 21:4205–4208

Zhang Z, Wang Y, Vikis HG et al (2001) Wildtype Kras2 can inhibit lung carcinogenesis in mice. Nat Genet 29:25–33