Cell. Mol. Life Sci. 64 (2007) 1752 – 1762
1420-682X/07/141752-11
DOI 10.1007/s00018-007-7054-6
© Birkhäuser Verlag, Basel, 2007

Cellular and Molecular Life Sciences

Cancer systems biology: exploring cancer-associated genes on cellular networks

E. Wang[a,]*, A. Lenferink[b], and M. O'Connor-McCourt[b]

[a] Computational Chemistry and Biology Group, Biotechnology Research Institute, National Research Council Canada, Montreal, Quebec, H4P 2R2 (Canada), Fax: 1-514-496-5143, e-mail: edwin.wang@cnrc-nrc.gc.ca
[b] Receptors, Signaling and Proteomics Group, Biotechnology Research Institute, National Research Council Canada, Montreal, Quebec, H4P 2R2 (Canada)

Online First 6 April 2007

Abstract. Genomic alterations lead to cancer complexity and form a major hurdle for comprehensive understanding of the molecular mechanisms underlying oncogenesis. In this review, we describe recent advances in studying cancer-associated genes from a systems biology point of view. The integration of known cancer genes onto protein and signaling networks reveals the characteristics of cancer genes within networks. This approach shows that cancer genes often function as network hub proteins which are involved in many cellular processes and form focal nodes in information exchange between many signaling pathways. Literature mining allows constructing gene-gene networks, in which new cancer genes can be identified. The gene expression profiles of cancer cells are used for reconstructing gene regulatory networks. By doing so, genes which are involved in the regulation of cancer progression can be picked up from these networks, after which their functions can be further confirmed in the laboratory.

Keywords. Cancer, systems biology, protein network, network biology, signaling network, cancer gene hunting, reverse engineering.

Introduction

Cancer is an extremely complex, heterogeneous disease, which displays a degree of complexity at the physiological, tissue and cellular levels. The interactions between tumors and their microenvironments reflect the physiological complexity of cancers, which is the recent focus of cancer research. Bidirectional interactions between cancer and its microenvironment might promote their growth, survival and the occurrence of distant metastasis [1]. However, the molecular mechanisms underlying the interactions between cancer cells and their microenvironment are poorly understood. A cancer tissue or a tumor often contains several distinct pathological cancer subtypes, which is recognized as cancer tissue complexity. This tissue complexity is believed to provide functional redundancy for tumors to maintain cellular heterogeneity, which could lead to tumor recurrence [2–4] as long as a cancer subtype or a fraction of cancer cells with metastatic potential survives after anticancer treatment. One cancer subtype is able to functionally replace another or even multiple subtypes killed by medical treatments such as anti-cancer drugs [5, 6]. The functional replacement of cancer subtypes allows tumor survival, further proliferation and finally tumor recurrence. It is reasonable to think that each cancer cell subtype within a tumor might originate through different cancer-specific developmental mechanisms and mutations in distinct genes. Therefore, this complexity will require a combination of several drugs or treatments targeting various cancer cell subtypes within a tumor. Work in the past few years, has identified the molecular signatures of various cancer subtypes in tumors through large-scale gene expression profiling analyses using microarray technology. For example, sets of gene expression signatures have

* Corresponding author.

been identified for breast cancer subtypes [7–9]. Nevertheless, for effective cancer treatment, it is necessary to identify those oncogenic signaling pathways that are the driving force for each of these cancer cell subtypes. Linking cancer subtypes to oncogenic signaling pathways and cascades is still hampered by poor understanding of the oncogenic processes at the cellular level. The coexistence of several cancer cell subtypes, which rely on activation of different signaling pathways in one particular tumor, represents the tissue complexity of cancers, while activation of multiple pathways that lead to the development of the same type of cancer represent the cellular complexity of cancers.

Cancer cells characteristically display uncontrolled cell growth, and the ability to invade surrounding tissue and finally to generate metastasis in distant places of human body. The accumulation of genetic mutations in part triggers tumor development and progression. Gene mutation or deregulation also promotes cell mobility that is highly correlated with tissue invasion and the formation of distant metastasis. In cancer, many kinds of gene alterations, such as gene sequence mutations [10, 11], gene and chromosomal fragment amplifications, chromosomal translocations and gene fusions [12–15], gene deletions [16, 17] and even the mutations and deregulations of noncoding RNAs, such as microRNAs [18–20], have been studied and documented extensively. A recent genome-wide screening of cancer mutation genes revealed that different cancer clinical samples of a same cancer type contain different sets of mutated genes which have divergent functions, indicating that the mutated genes do not belong to same pathway, and therefore suggesting that a cancer could develop through multiple genetic routes [21]. Because gene activity and regulation ultimately define a cancer phenotype, it is essential to have a comprehensive understanding of the precise genetic mutations and consequences of these mutations and genetic alterations. Therefore, it is not surprising that the majority of research efforts focus on the genomics, functional genomics and proteomics of cancer cell progression and metastasis.

The complexity of cancer is a major obstacle to comprehensive understanding of the underlying molecular mechanisms of oncogenesis. No gene is an island. Even in a single cell, genes work together and take part in many biological processes which then determine the cell's behavior and phenotype. Scientists have struggled for years to figure out how to handle this biological complexity. Systems biology, or more specifically network biology, is driven by the gradual realization that a particular biological function is not the result of activity encoded by a single gene. The goal of systems biology is to combine molecular information of various types in models to understand biological systems and their complexity, and finally to attempt to predict biological function at the cellular, tissue, organ and even whole-organism level. Development of genomic technologies such as high-throughput sequencing, especially DNA, protein microarrays and mass spectrometry, has made it possible to describe cells' biological states in a quantitative manner, and to simultaneously study many gene and protein components and then clarify how these components work together in regulation and carrying out biological processes. The integration of these experimental techniques with information technology provides a powerful approach to address and dissect the complexity of cancer and other biological problems at various levels in a systems manner.

Biological understanding of cellular networks

In cells, interdependent interactions of genes and proteins form complex cellular networks, for example signaling networks, gene regulatory networks and metabolic networks. Cellular networks are the basis of biological complexity. Therefore, the cellular networks have thus become the core of systems biology. Traditionally, network and graph theory is a branch of mathematics. Here we briefly review and explain network and graph theory with a focus on biological insights. Recent developments in high-throughput techniques in the field of genomics and proteomics research have generated vast amounts of data; furthermore, information in the literature is becoming accessible on the Internet. Extraction of these datasets and information to generate new cellular networks or to integrate into and expand existing cellular networks makes it attractive to study the structures of these networks by relating them to biological properties and insights. What is needed now it to develop systematic methods for analyzing cellular networks as well as understanding their properties in a cellular context.

In biology, cellular networks include protein interaction networks which encode the information of proteins and their physical interactions, signaling networks which illustrate inter- and intracellular communication and the information process between signaling proteins, gene regulatory networks which describe regulatory relationships between transcription factors and/or regulatory RNAs and genes, and metabolic networks of biochemical reactions between metabolic substrates and products. Metabolic networks are not the focus of this review; however, more information about metabolic networks can be found in a recent review [22]. Subcellular networks include amino acid residue interaction networks in protein structures, which are assumed to involve a permanent flow of information between amino acids [23].

Networks can be presented as either directed or undirected graphs. Protein interaction networks are modeled as undirected graphs, in which the nodes represent proteins and the links represent the physical interactions between the proteins. Directed graphs, on the other hand, are used to present gene regulatory and metabolic networks. In gene regulatory networks, nodes represent transcription factors or genes, while links represent regulatory relations between transcription factors and the regulated genes or transcription factors. Signaling networks are presented as graphs containing both directed and undirected links. In the networks nodes represent proteins, directed links represent activation or inactivation relationships between proteins, while undirected links represent physical interactions between proteins. Compared with other cellular networks, signaling networks are far more complex in terms of the relationships between proteins. For example, nodes may represent different functional proteins, such as kinases, growth factors, ligands, receptors, adaptors, scaffolds, transcription factors and so on, which all have different biochemical functions and are involved in many different types of biochemical reactions that characterize a specific signal transduction machinery.

In the past few years, significant progress has been made in the identification and interpretation of the structural properties of cellular networks. This information has shed light on how such properties might reflect the biological meaning and behavior of cellular networks [24, 25]. Although each type of the cellular network has its own properties, they all share some common structural properties. Cellular networks and other real-world networks, such as a public transportation network, exhibit a global structure property that is defined as 'scale-free'. In a scale-free network, a small group of nodes act as highly connected hubs (high degree), whereas most nodes have only a few links (low degree). For example, a map describing the air transportation in the United States is a network, in which only a few big airports (hubs) in big cities such as Boston, New York, Chicago and Los Angeles have many air routes (links) to other airports, while many small airports have just a few air routes to big airports nearby. This common structural feature encodes a special property of these networks: they are robust but also very vulnerable to failure and attack [25]. In a scale-free network, random removal of a substantial fraction of the low-degree nodes will cause little damage to the network's connectivity; however, targeted removal of the high-degree hub nodes will easily disconnect and destroy the network completely, as illustrated by the air transportation map. Disabling big airports (hubs) will wreak havoc in many ways, while damaging a few small airports will have little or no effect on overall air transportation.

In regulatory networks, hub genes are global transcription factors. They may govern a large number of genes in response to internal and external signals. To fit their multiple biological functions, the hub's expression will have to display dynamic features. Analysis of the yeast gene regulatory network, in which the gene expression profiles of many different cellular conditions were integrated, shows that the hub transcription factors do control a large spectrum of biological processes [26]. We have integrated genome-wide messenger RNA (mRNA) decay data onto the *Escherichia coli* gene regulatory network and revealed that the transcription factors whose mRNAs have fast decay rates are significantly enriched in hub genes, suggesting that the expression of the hub genes in gene regulatory networks are indeed highly dynamic. This dynamic behavior facilitates a rapid response of the network to external stimuli [27]. A similar result was obtained in a recent study, in which mRNA decay data mapped onto a yeast protein interaction network showed that the hub proteins in protein interaction networks also display fast mRNA decay rates [28]. In protein interaction networks, hub proteins are involved in a large number of interactions, meaning that these proteins will take part in many biological processes and therefore would have higher dynamics in expression. Furthermore, hub proteins may be more important for an organism's survival and have a much broader effect on a system than non-hub proteins. A series of reports confirm this notion [24, 29–32]. These reports also suggest that hub proteins have central positions in cellular networks and are more essential for the organism's survival than other proteins. Therefore, the structure, or in other words, the topology, of cellular networks not only sheds light on complex cellular mechanisms and processes but also gives insight into evolutionary aspects of the proteins involved. By examining protein evolution and protein interaction networks, Saeed and Deane found that hub proteins are 'old' proteins which have evolved more slowly than other proteins [33]. Biologically, this makes perfect sense, in that hub proteins are involved in many biological processes and are subject to selection pressure and constraints. Hub proteins in signaling networks are the focal nodes that are shared by many signaling pathways. That is, hub proteins have become information exchanging and processing centers. Alterations to these hub proteins may therefore globally affect the well-being of living cells. A recent RNA interference (RNAi) screen of worms supports this hypothesis. Lehner et al. systematically mapped the genetic interactions of *Caenorhabditis elegans* genes involved in signaling pathways and revealed a network of 350 interactions [34]. They then tested 65,000 pairwise gene interactions and found that a few genes interact with an unexpectedly large number of signaling pathways. These hub genes were identified as chromatin-modifying proteins, which are conserved across animals where they display core genetic buffering properties.

Cellular networks are complex systems in which a gene does not independently perform a single task. Instead, individual genes, which collaborate to carry out some specific biological function can be grouped. We call such a gene group a functional module. This assumption leads to the idea that a complex network can be broken up into many small but functional modules or units, which can then be studied to determine their structural properties and functional behaviors. Once we understand the functions, properties and regulatory/interactive behaviors of these modules, we can then use these functional modules to rebuild subnetworks and even whole networks and study their properties and functions. Network motifs are examples of such functional modules. These are statistically significant recurring structural patterns or small subgraphs or subnetworks that are found more often in a real network than would be expected by chance [35]. These motifs are known in biology as gene regulatory loops. These motifs can self-organize or form a networks by sharing nodes between various motifs [27]. Network motifs have been studied in detail in gene regulatory networks. Three major motifs are found in gene regulatory networks: single input module (SIM), bi-fan and feedforward loop (FFL) (Fig. 1). One design principle of these motifs is that the transcription factors whose mRNAs have fast decay rates are significantly enriched, suggesting that motif structures encode regulatory behavior: network motifs are able to rapidly respond to internal and external stimuli and decrease internal cell noise [27]. Network motifs have been shown to have distinct regulatory functions and are robust in that they are resistant to internal noise. Both theoretical and experimental studies have shown that network motifs have distinct regulatory functions and particular kinetic properties that determine the temporal program of gene expression [36]. Therefore, the frequencies and types of network motifs cells use reveal the regulatory strategies that are selected in different cellular conditions [27, 37, 38]. For example, FFLs are buffers that respond only to persistent input signals [39], which makes them wellsuited for responding to endogenous conditions, while the motifs whose key regulatory transcripts have a fast mRNA decay rate are preferentially used for responding to extraneous conditions [27]. In signaling networks, network motifs such as switches [40], gates [41], and positive or negative feedback loops provide specific regulatory capacities in decoding signal strength, processing information and controlling noise [42, 43].

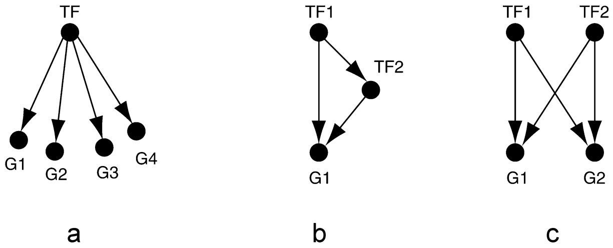

Figure 1. Network motifs in gene regulatory networks. Nodes represent genes and lines represent gene regulatory relations. (*a*) Single input module (SIM): a transcription factor (TF) regulates a group of genes (G1, G2, G3 and G4). (*b*) Feedforward loop (FFL): a transcription factor (TF1) regulates the second transcription factor (TF2); both TF1 and TF2 regulate a target gene (G1). (*c*) Bi-fan: both transcription factors, TF1 and TF2, regulate both target genes (G1 and G2).

Distinct network motifs could form large aggregated structures, called network themes, that perform specific functions by forming collaborations among a large number of motifs [44]. In this case, network themes can be regarded as communities of functionally related nodes. A large protein complex in protein interaction networks is one of the examples of such a network community.

Integrative network analysis of cancer-associated genes

High-throughput gene expression profiling often leads to identification of hundreds or sometimes even thousands of modulated genes for a given phenotype. However, the extraction and interpretation of biological insights of the differentially expressed genes in these high-throughput datasets are challenging, and limited by difficulties in recognizing gene-gene relations and associations within a huge amount of data. Although it is possible to classify identified genes into different functional groups using gene ontology (GO) [45], the in-depth relationships between genes in different functional categories can still not be easily illustrated. A particular phenotype is the result of collaborations of a group of genes, which do not necessarily belong to the same functional category. Therefore, integration of microarray-generated gene lists into cellular networks could help in analyzing and interpreting the biological significance of the genes in a network and their gene-interdependent context. This notion provides a structured network knowledge-based approach to analyze genome-wide gene expression profiles in the context of known functional interrelationships among genes, proteins and phenotypes.

Motivated by this concept, Wachi et al. investigated differentially expressed genes in squamous cell lung cancer which were identified by projecting microarray gene expression profiling onto a human protein interaction network [46]. The data for the network construction were taken from the online predicted human interaction database, (OPHID) [47], which contains 16,034 known human protein interactions obtained from various public protein interaction databases, and 23,889 additional protein interactions that were predicted. They mapped the 360 upregulated and 270 downregulated genes that were identified in the lung cancer microarray experiment onto the protein interaction network. Further network analysis revealed that the upregulated genes in this dataset are well connected, whereas the suppressed genes and randomly selected genes are less so. They also showed a high degree of centrality in these differentially upregulated genes, but not for the genes that are suppressed. These results imply that the upregulated, but not downregulated genes in this experiment are enriched in hub proteins, which are associated with essential functions in protein interaction networks [29]. Cancer cells are characterized by uncontrolled growth, which could suggest that the induced genes in cancer cells, compared with normal cells, are more essential for survival and proliferation. The work described here reveals the characteristics of cancer-associated genes in a network context and supports the notion that integrative network analysis of large datasets obtained from gene expression profiling helps to understand the function of biological systems.

The characteristics of cancer-associated genes uncovered in this study were confirmed by a recent analysis of a human protein interaction network integrated with literature-mined cancer genes. Johsson and Bates [48] used mutated cancer genes collected from the literature [49] and attempted to uncover their intrinsic properties in a human protein interaction network which was constructed from the entire human genome using an orthology-based method [50]. In total, 346 genes encoding 509 protein isoforms were mapped onto the network. This analysis showed that cancer proteins have on average twice as many interaction partners as other proteins in the network, which implies the evolutionary aspects of cancer genes. Accumulating evidence shows a positive correlation between the evolution of proteins and their number of interactions within a given network [31, 51, 52]. With this consideration in mind, the authors concluded that proteins whose mutation results in a detrimental change of function that leads to cancer may generally be more conserved than other proteins. Alternatively, as they have more interaction partners, cancer proteins, may be involved in significantly more biological processes and play a central role in the protein network. To further explore this direction, Johsson and Bates also investigated the relationships between these cancer genes and network communities, which represents a distinct biological process, meaning that if a protein is a member of multiple network communities, it takes part in more biological processes. The results of this analysis show that the identified cancer proteins are indeed involved in more network communities than other proteins in the network, suggesting their more prominent centrality and participation in the formation of the proteome network backbone. Taking it one step further, the authors also analyzed the domain compositions of these cancer proteins. Cancer proteins display a high ratio of highly promiscuous domains in terms of the number of different proteins with which they interact, indicating that they play central roles in many biological processes and that mutations in these proteins could lead to a higher cancer incidence. Moreover, the domains most frequently found in the cancer protein population have functionalities that particularly focus on DNA regulation and repairing, such as Zinc-finger, PHD-finger, BRCT and Paired-box domains, which all happen to be transcription factors.

These findings provide biological insight into the global protein interaction network properties of cancer proteins and uncover one of the most striking properties of cancer proteins in that cancer-associated proteins are network hubs, which play central roles in biological systems and take part in many biological processes. Taken together, each hub cancer protein may reflect a specific domain of a cellular function, which suggests that mutations of an individual or a few hub proteins together may lead to oncogenesis or cancer progression. However, these studies provide little insight into the oncogenic mechanisms simply because protein interaction networks have limited information compared with signaling networks in which protein regulatory (activation and blocking) information is encoded. Therefore, integration of cancer genes into existing and established signaling networks would enable further insight into the oncogenic process and cancer progression.

Cells use sophisticated communication between proteins to perform a series of tasks such as growth and maintenance, cell survival, apoptosis and development. Signaling pathways are crucial to maintain cellular homeostasis and determine cell behavior. Therefore, alterations in the expression of genes and their regulators will be reflected in these cellular signaling pathways, and in turn lead to tumor development and/or the promotion of cell migration and metastasis. Indeed, mutations in genes that encode signaling proteins are commonly observed in many types of cancers [53].

Specific signaling pathways deploy many different proteins; however, pathways often 'talk' each other. This so-called 'cross-talk' between pathways has been systematically investigated in a recent study, and an unexpectedly high number of cross-talk events among signaling pathways were discovered [54]. These results indicate that signaling pathways form a complex network to process information. Structural analysis of a literature-mined human cellular signaling network containing ~500 proteins showed that signaling pathways are intertwined in order to manage the numerous cell behavior outputs [55]. This work provides a framework for our understanding of how signaling information is processed in cells. Furthermore, analysis of interactions between microRNAs and the same signaling network reveals the principles of microRNA regulation of the network [56]. Together, these approaches suggest that an integrative analysis of signaling networks with cancer proteins would highlight the characteristics of cancer proteins within these networks.

Errors in signal transduction can lead to altered development and incorrect behavioral decisions, which could result in uncontrolled cell growth or even cancer. The relationships of signaling proteins are thought to be critical in determining cell behavior. Therefore, mapping of cancer genes on the nodes of a signaling network could in general, lead us to which mechanisms support the continued survival and proliferation of cancer cells. We manually curated human cellular signaling pathways and merged these curated data into another literature-mined human cellular signaling network mentioned previously [55]. As a result, the new network contains ~1.100 signaling proteins. Next the cancer proteins obtained from NCBI's Online Mendelian Inheritance in Man (OMIM) database [57] were mapped onto the network. Nearly 90 cancer proteins were mapped onto the network [58]. Not surprisingly, cancer proteins are enriched in hub proteins in the signaling network. As mentioned, cancer genes often become mutated, which could result in the activation of particular focal signaling nodes that play important roles in the information exchange between many individual signaling pathways. Indeed, several cancer proteins form the focal nodes in signaling networks and therefore play important roles in cancer development.

The cellular signal information flow initiates from the extracellular space, e.g. a ligand binds to a cellular membrane receptor to generate the signal that is then transmitted by intracellular signaling components in the cytosol to the signaling components within the nucleus. This process of signal transduction is sensitive in terms of mutated genes, which result in altered signaling and therefore tumorigenesis, and increase cell mobility and invasion. We found that cancer proteins are enriched in the downstream section of signaling networks, the realm of the transcription factors [58]. Along with this discovery, we also found that cancer proteins are hardly represented in certain network motifs, such as bi-fan (Fig. 1), which is a structure with regulatory redundancy and also one of the most abundant network motifs in the central region of the human signaling network. These results lead us to believe that the central region of a signaling network provides a genetic buffer for cells in that it may prevent cancer development, which is in agreement with the robustness of networks [59]. The fact that cancer proteins are enriched in the downstream region suggests that proteins in this region are crucial for determining specific cell behavior. Our work provides insights into the signaling networks invoked in cancer development and progression.

The systems-level approach taken in this work, i.e. combining information on how proteins interact with each other and how transmitted signals are processed, with information on known cancer genes and gene expression in cancer cells, is a particularly appealing approach to gain an understanding of complex biological processes, such as cancer development and metastasis. Network analyses using comprehensive knowledge of biology provide a framework for structuring existing knowledge regarding cancer biology and help to identify proteins and/or significant functional modules and the underlying mechanisms of the oncogenic process.

Hunting new cancer-related genes using cellular networks

Protein interaction networks have been used to hunt new cancer-associated genes. Jonsson et al. have been motivated to find genes involved in metastasis by integrating cancer cell microarray expression data onto a rat protein interaction network which was constructed by transferring protein-protein interaction information from other species using the protein homology concept [50]. The network was evaluated by confidence scores based on homology to proteins that have been experimentally observed to interact. Metastasis is a key event that is usually associated with a poor prognosis in cancer patients. Metastasizing cancer cells have special properties, in that they can display features such as increased motility and invasiveness.

It was hypothesized that subnetworks of protein interactions may govern metastasis. Jonsson et al. used a dataset containing up- and downregulated genes that was obtained from a cancer microarray study, and constructed subnetworks around proteins which were then evaluated using cluster analysis to define network communities that reflect small protein interaction units that are involved in the metastastic process [50]. As a result, they identified 37 protein communities of highly interconnected proteins, most of which have been associated with cancer and metastasis.

Gene networks have been constructed by merging various data sources, which were then used to find or prioritize cancer and other disease genes. In this context, gene-gene networks are presented using undirected graphs, in which the nodes represent genes and the links represent relations between genes. The relations of the genes can be physical protein interactions, gene regulatory relations, gene associations and so on. Franke et al. constructed such a human gene-gene network using databases of known interactions, gene ontology (GO), microarray co-expressions and yeast two-hybrid data [60]. They then integrated this network with already known genetic information on diseases (i.e. genetic loci for a particular human disease). The authors reasoned that the cancer genes from each locus are likely to be involved in one same molecular pathway and biological process. To prove their concept, they showed that the genes prominent in any one disease were closer to each other in the network than would be expected by chance, which suggests that these genes are involved in the disease and therefore tend to have more functional interactions or associations. To assess the predictive power of this method, the authors tested it by picking disease genes using the network. Four out of 10 breast cancer genes were ranked at the top of the gene list, which is 4 times higher than a breast cancer gene that would be picked by chance. When they integrated more interaction data into the network and adjusted the network topology, the ranking of these disease genes improved considerably, and included 9 of the 10 genes. These results indicate that the use of a network significantly improves the chance of finding the correct cancer genes.

In the past few years, a series of studies focused on constructing gene-gene networks using data from literature and other sources. One notion behind this is that nearly 80 % of biological information and data is coded in natural language in technical reports, websites, research publications and other text documents [61]. To facilitate the extraction of these data, methods have been developed for the automatic extraction of interaction and pathway information from the scientific literature [62–66]. Furthermore, the extracted relations between genes have been used to construct gene-gene networks, and several software packages and related datasets have been developed. PubGene [67] is an example of such a tool. It contains a database and analysis software for constructing gene-gene networks by identifying relationships between genes based on their statistical co-occurrence in the abstracts of scientific papers. Information Hyperlinked over Proteins (iHop) [68] is another example. In this case, one can use gene names to retrieve gene-gene relations from PubMed abstracts that match a specified gene/protein name. iHop also provides automatic extraction gene-gene relations for software developers and bioinformatics scientists.

In contrast to most text-mining methods that use the abstracts of research papers, Natarajan et al. tried to use full-length scientific articles to extract gene-gene relations [69], and also fused the extracted gene interactions to structured data and knowledge bases such as Ingenuity Pathway Analysis, UniProt [70], InterPro [71], NCBI Entrez and GO. A human gene-gene network was constructed using theses data sources. The authors then mapped the differentially expressed genes identified from microarrays, which profiled the gene expression in glioblastoma as a response to S1P in vitro. Further analysis led to identification of a cascading event that is triggered by S1P, and which leads to the transactivation of MMP-9 via neuregulin-1, vascular endothelial growth factor and the urokinase-type plasminogen activator. This suggests that the interaction network has the potential to shed new light on our understanding of the cancer-related process. Therefore, automated extraction of information from the biological literature, together with combining and integrating biological data from laboratory experiments, provides an effective approach to biological knowledge discovery.

Reverse engineering of gene regulatory networks from microarray data

Reverse engineering of biological networks is a process of elucidating the structure of gene regulation relationships by reasoning backwards from the observations of gene expression values. In recent years, a substantial effort has been made to reconstruct gene regulatory networks using microarray profiles. Here we describe two related efforts which combined computational and experimental approaches.

Basso et al. developed a statistical algorithm using mutual information for more accurately reasoning

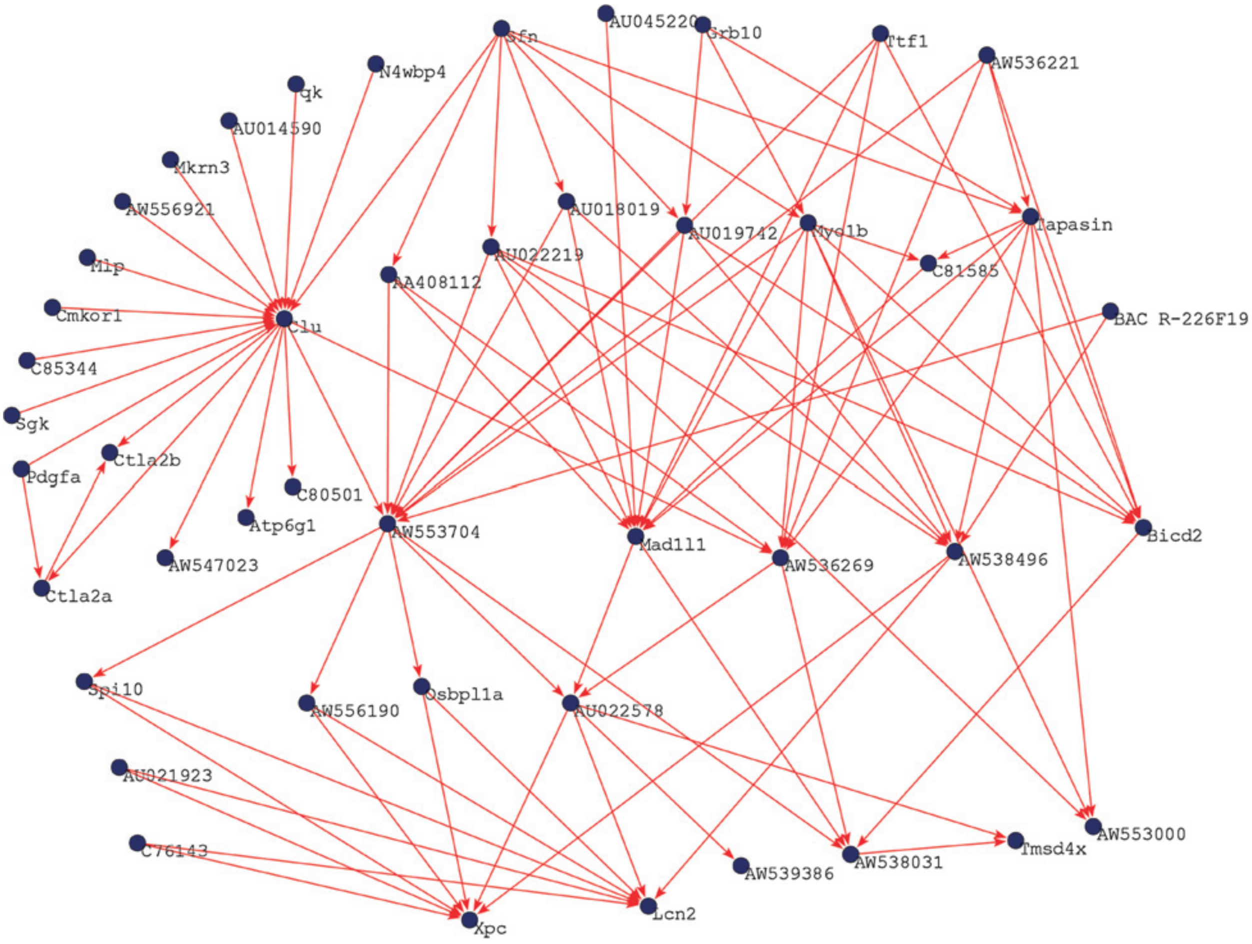

Figure 2. A gene regulatory network inferred from the time course gene expression profiles of the BRI-JM01 cell line. Nodes represent genes, and lines represent gene regulatory relations.

networks in which pairwise gene-gene interactions are described [72]. The algorithm was named the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE). To test ARACE, the authors used a huge number of gene expression profiles (336 samples) of human B-cells at different stages covering normal to cancer cells to construct a network. A subnetwork was used for validation using GO and chip-on-chip experiments. The results are encouraging in that 90 % specificity was obtained for ARACNE. However, note that the test did not include the predictions with lowest mutual information scores. Nevertheless, this approach shows that with enough gene expression data, reasonable gene networks can be retrieved by developing proper algorithms.

Another example of reverse engineering applied to cancer research was carried out using a dataset that was generated in our own laboratory. We constructed a gene regulatory network using time course microarray profiles from a mouse epithelial breast cell line (BRI-JM01), which was isolated from mammary tumors in transgenic mice. These cells undergo an epithelial to mesenchymal transition (EMT) when they are treated with TGF-β (transforming growth factor β) [73]. To identify the transcriptional changes underlying this EMT, we exposed the BRI-JM01 cell line to TGF-β for seven time intervals (0.5 – 24 h), and interrogated the transcriptome using complementary DNA (cDNA) microarrays. Based on the microarray profiles and the Markov chain-based network construction method [74], we constructed a gene regulatory network that contains nearly 50 genes and three layers of regulations, in which the regulatory relations are either direct or indirect (Lenferink et al., unpublished data, Fig. 2). Known biological information was used to validate the network. Interestingly, in the top layer of the network, all the annotated genes are either transcription factors or signaling proteins which are known to be regulatory proteins. Most known genes in the bottom layer of the network are involved in cancer processes, which suggests that the network may be right. Notably, clusterin, one of the genes that are upregulated in the middle and late time points shows many regulatory links to other genes in the network. During the EMT process, clustrerin is secreted by BRI-JM01 cells. Interestingly, when applying anti-clusterin antibodies to TGF-β-treated BRI-JM01 cells, we were able to block TGF-β-induced EMT. This result strongly implies that the secreted form of clusterin plays a pivotal role in TGF-β-induced EMT and therefore TGF-β's tumor-promoting effects on the BRI-JM01 cell line. Currently, reverse engineering of gene regulatory networks using microarray data is mainly hampered by limited microarray experiments we could perform for a given sample. Reverse engineering methods provide only some hints to biologists, although they could narrow down the gene list of interest. Substantial lab experiments should be followed to further validate genes of interest from the inferred gene regulatory networks.

Outlook

The analysis of the cancer phenomenon using a systems-level approach is still in its infancy. New and emerging technologies need to be developed and validated. These technologies include single-cell signal mapping, which will be very helpful in obtaining the full picture of signaling dynamics occurring in different cancer cells and during various stages of cancer development. These techniques will be especially useful for understanding the biology of tumors, which consist of notoriously heterogeneous cancer cell populations. Information about relations between genes and/or proteins is still limited, but will be alleviated once new high-throughput datasets become available. These new datasets, whether generated experimentally or by literature mining, will no doubt provide information on new interactions between genes. Current efforts are ongoing to curate high-quality signaling data from the literature [75, 76].

Overall, the systems biology output will bring unprecedented amounts of molecular information and large-scale datasets to medicine in the form of DNA sequences and quantative information on messenger RNS, proteins, and metabolites. An important part of systems biology is to take all of these measurements in consideration to construct models to describe what is going on in a cell, a tissue, an organ or even an organism. A systems-level understanding of the underlying mechanisms causing cancer in an individual cancer patient will allow science to become more focused and will contribute significantly to the clinical application of personalized medicine.

1 Alberti, C. (2006) Prostate cancer progression and surrounding microenvironment. Int. J. Biol. Markers 21, 88 – 95.

2 Harris, J. F., Chambers, A. F., Hill, R. P. and Ling, V. (1982) Metastatic variants are generated spontaneously at a high rate in mouse KHT tumor. Proc. Natl. Acad. Sci. USA 79, 5547 – 5551.

3 Ling, V., Chambers, A. F., Harris, J. F. and Hill, R. P. (1984) Dynamic heterogeneity and metastasis. J. Cell Physiol. Suppl. 3, 99 – 103.

4 Chambers, A. F., Harris, J. F., Ling, V. and Hill, R. P. (1984) Rapid phenotype variation in cells derived from lung metastases of KHT fibrosarcoma. Invasion Metastasis 4, 225 – 237.

5 Bissell, M. J. and Radisky, D. (2001) Putting tumours in context. Nat. Rev. Cancer 1, 46 – 54.

6 Petersen, O. W., Lind, N. H., Gudjonsson, T., Villadsen, R., Ronnov-Jessen, L. and Bissell, M. J. (2001) The plasticity of human breast carcinoma cells is more than epithelial to mesenchymal conversion. Breast Cancer Res. 3, 213 – 217.

7 Kapp, A. V., Jeffrey, S. S., Langerod, A., Borresen-Dale, A. L., Han, W., Noh, D. Y., Bukholm, I. R., Nicolau, M., Brown, P. O. and Tibshirani, R. (2006) Discovery and validation of breast cancer subtypes. BMC Genomics 7, 231.

8 Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R. R. and Borresen-Dale, A. L. (2006) Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics 7, 127.

9 Bertucci, F., Finetti, P., Rougemont, J., Charafe-Jauffret, E., Cervera, N., Tarpin, C., Nguyen, C., Xerri, L., Houlgatte, R., Jacquemier, J., Viens et al. (2005) Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. Cancer Res. 65, 2170 – 2178.

10 Blons, H., Cote, J. F., Le, C. D., Riquet, M., Fabre-Guilevin, E., Laurent-Puig, P. and Danel, C. (2006) Epidermal growth factor receptor mutation in lung cancer are linked to bronchioloalveolar differentiation. Am. J. Surg. Pathol. 30, 1309 – 1315.

11 Lievre, A., Bachet, J. B., Le, C. D., Boige, V., Landi, B., Emile, J. F., Cote, J. F., Tomasic, G., Penna, C., Ducreux, M. et al. (2006) KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res. 66, 3992 – 3995.

12 Zhang, X. Y., Hu, Y., Cui, Y. P., Miao, X. P., Tian, F., Xia, Y. J., Wu, Y. Q. and Liu, X. (2006) Integrated genome-wide gene expression map and high-resolution analysis of aberrant chromosomal regions in squamous cell lung cancer. FEBS Lett. 580, 2774 – 2778.

13 Hughes, S., Yoshimoto, M., Beheshti, B., Houlston, R. S., Squire, J. A. and Evans, A. (2006) The use of whole genome amplification to study chromosomal changes in prostate cancer: insights into genome-wide signature of preneoplasia associated with cancer progression. BMC Genomics 7, 65.

14 Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R. F., Barany, F.et al. (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res. 66, 2129 – 2137.

15 Taki, T. and Taniwaki, M. (2006) Chromosomal translocations in cancer and their relevance for therapy. Curr. Opin. Oncol. 18, 62 – 68.

16 Cesar, A. C., Borim, A. A., Caetano, A., Cury, P. M. and Silva, A. E. (2004) Aneuploidies, deletion, and overexpression of TP53 gene in intestinal metaplasia of patients without gastric cancer. Cancer Genet. Cytogenet. 153, 127 – 132.

17 Gallegos-Arreola, M. P., Gomez-Meda, B. C., Morgan-Villela, G., Arechavaleta-Granell, M. R., Arnaud-Lopez, L., Beltran-Jaramillo, T. J., Gaxiola, R. and Zuniga-Gonzalez, G. (2003) GSTT1 gene deletion is associated with lung cancer in Mexican patients. Dis. Markers 19, 259 – 261.

18 Calin, G. A. and Croce, C. M. (2006) MicroRNA-cancer connection: the beginning of a new tale. Cancer Res. 66, 7390 – 7394.

19 Volinia, S., Calin, G. A., Liu, C. G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M. et al. (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. Proc. Natl. Acad. Sci. USA 103, 2257 – 2261.

20 Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M.et al. (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 65, 7065 – 7070.

21 Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N. et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314, 268 – 274.

22 Sweetlove, L. J. and Fernie, A. R. (2005) Regulation of metabolic networks: understanding metabolic complexity in the systems biology era. New Phytol. 168, 9 – 24.

23 del, Sol, A., Fujihashi, H., Amoros, D. and Nussinov, R. (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol. Syst. Biol. 2, 2006.0019.

24 Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283 – 291.

25 Barabasi, A. L. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101 – 113.

26 Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308 – 312.

27 Wang, E. and Purisima, E. (2005) Network motifs are enriched with transcription factors whose transcripts have short half-lives. Trends Genet. 21, 492 – 495.

28 Batada, N. N., Hurst, L. D. and Tyers, M. (2006) Evolutionary and physiological importance of hub proteins. PLoS Comput. Biol. 2, e88.

29 Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. Nature 411, 41 – 42.

30 Wuchty, S., Oltvai, Z. N. and Barabasi, A. L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet. 35, 176 – 179.

31 Wuchty, S. (2004) Evolution and topology in the yeast protein interaction network. Genome Res. 14, 1310 – 1314.

32 Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K. et al. (2005) A network-based analysis of systemic inflammation in humans. Nature 437, 1032 – 1037.

33 Saeed, R. and Deane, C. M. (2006) Protein protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics 7, 128.

34 Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A. G. (2006) Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. Nat. Genet. 38, 896 – 903.

35 Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli.* Nat. Genet. 31, 64 – 68.

36 Mangan, S., Zaslaver, A. and Alon, U. (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J. Mol. Biol. 334, 197 – 204.

37 Balazsi, G., Barabasi, A. L. and Oltvai, Z. N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli.* Proc. Natl. Acad. Sci. USA 102, 7841 – 7846.

38 Kalir, S., Mangan, S. and Alon, U. (2005) A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli.* Mol. Syst. Biol. 1, 2005.0006.

39 Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100, 11980 – 11985.

40 Bhalla, U. S., Ram, P. T. and Iyengar, R. (2002) MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. Science 297, 1018 – 1023.

41 Blitzer, R. D., Connor, J. H., Brown, G. P., Wong, T., Shenolikar, S., Iyengar, R. and Landau, E. M. (1998) Gating of CaMKII by cAMP-regulated protein phosphatase activity during LTP. Science 280, 1940 – 1943.

42 Angeli, D., Ferrell, J. E., Jr. and Sontag, E. D. (2004) Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Natl. Acad. Sci. USA 101, 1822 – 1827.

43 Dublanche, Y., Michalodimitrakis, K., Kummerer, N., Foglierini, M. and Serrano, L. (2006) Noise in transcription negative feedback loops: simulation and experimental analysis. Mol. Syst. Biol. 2, 41.

44 Zhang, L. V., King, O. D., Wong, S. L., Goldberg, D. S., Tong, A. H., Lesage, G., Andrews, B., Bussey, H., Boone, C. and Roth, F. P. (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. J. Biol. 4, 6.

45 Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 32, D262-D266.

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

47 Brown, K. R. and Jurisica, I. (2005) Online predicted human interaction database. Bioinformatics 21, 2076 – 2082.

48 Jonsson, P. F. and Bates, P. A. (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22, 2291 – 2297.

49 Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. R. (2004) A census of human cancer genes. Nat. Rev. Cancer 4, 177 – 183.

50 Jonsson, P., Cavanna, T., Zicha, D. and Bates, P. (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7, 2.

51 Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. and Feldman, M. W. (2002) Evolutionary rate in the protein interaction network. Science 296, 750 – 752.

52 Saeed, R. and Deane, C. M. (2006) Protein protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics 7, 128.

53 Rodriguez-Viciana, P., Tetsu, O., Oda, K., Okada, J., Rauen, K. and McCormick, F. (2005) Cancer targets in the Ras pathway. Cold Spring Harb. Symp. Quant. Biol 70, 461 – 467.

54 Natarajan, M., Lin, K. M., Hsueh, R. C., Sternweis, P. C. and Ranganathan, R. (2006) A global analysis of cross-talk in a mammalian cellular signaling network. Nat. Cell Biol. 8, 571 – 580.

55 Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J. et al. (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. Science 309, 1078 – 1083.

56 Cui, Q., Yu, Z., Purisima, E. O. and Wang, E. (2006) Principles of microRNA regulation of a human cellular signaling network. Mol. Syst. Biol. 2, 46.

57 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33, D514-D517.

58 Awan, A., Bari, H., Yan, F., Mokin, S., Yang, S., Chowdhury, S., Cui, Q., Yu, Z., Zhou, J., Purisima, E. and Wang, E. (2007) Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signaling network. IET Systems Biology, in press.

59 Kitano, H. and Oda, K. (2006) Robustness trade-offs and host-microbial symbiosis in the immune system. Mol. Syst. Biol. 2, 2006.0022.

60 Franke, L., Bakel, H. v., Fokkens, L., de Jong, E. D., Egmont-Petersen, M. and Wijmenga, C. (2006) Reconstruction of a functional human gene network with an application for prioritizing positional candidate genes. Am. J. Hum. Genet. 78, 1011 – 1025.

61 Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van, B., Jr. and Bremer, E. G. (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics 7, 373.

62 van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. and Leunissen, J. A. (2006) A text-mining analysis of the human phenome. Eur.. J. Hum. Genet. 14, 535 – 542.

63 Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. and Valencia, A. (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. Sci. STKE. May 10; 2005 (283), e21.

64 Bajdik, C. D., Kuo, B., Rusaw, S., Jones, S. and Brooks-Wilson, A. (2005) CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinformatics 6, 78.

65 McCallum, J. and Ganesh, S. (2003) Text mining of DNA sequence homology searches. Appl. Bioinformatics 2, S59-S63.

66 Eskin, E. and Agichtein, E. (2004) Combining text mining and sequence analysis to discover protein functional regions. Pac. Symp. Biocomput. 288 – 299.

67 Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28, 21 – 28.

68 Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21 Suppl. 2, ii252-ii258.

69 Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van, B., Jr. and Bremer, E. G. (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics 7, 373.

70 Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 34, D187-D191.

71 Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. et al. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res. 13, 662 – 672.

72 Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., la-Favera, R. and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382 – 390.

73 Lenferink, A. E., Magoon, J., Cantin, C. and O'Connor-McCourt, M. D. (2004) Investigation of three new mouse mammary tumor cell lines as models for transforming growth factor (TGF)-beta and Neu pathway signaling studies: identification of a novel model for TGF-beta-induced epithelial-to-mesenchymal transition. Breast Cancer Res. 6, R514-R530.

74 Dewey, T. G. and Galas, D. J. (2001) Dynamic models of gene expression and classification. Funct. Integr. Genomics 1, 269 – 278.

75 Oda, K., Matsuoka, Y., Funahashi, A. and Kitano, H. (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol. Syst. Biol. 1, 2005.0010.

76 Oda, K. and Kitano, H. (2006) A comprehensive map of the toll-like receptor signaling network. Mol. Syst. Biol. 2, 2006.0015.