Comparative Analysis of Protein-Protein Interactions in Cancer-Associated Genes

Purnima Guda[1], Sridar V. Chittur[2], and Chittibabu Guda[1]*

[1] *GenNYsis Center for Excellence in Cancer Genomics and Department of Epidemiology & Biostatistics, State University of New York at Albany, Rensselaer, NY 12144-3456, USA;* [2] *Center for Functional Genomics and Department of Biomedical Sciences, State University of New York at Albany, Rensselaer, NY 12144-3456, USA.*

*Corresponding author. E-mail: cguda@albany.edu
DOI: 10.1016/S1672-0229(08)60030-3

Protein-protein interactions (PPIs) have been widely studied to understand the biological processes or molecular functions associated with different disease systems like cancer. While focused studies on individual cancers have generated valuable information, global and comparative analysis of datasets from different cancer types has not been done. In this work, we carried out bioinformatic analysis of PPIs corresponding to differentially expressed genes from microarrays of various tumor tissues (belonging to bladder, colon, kidney and thyroid cancers) and compared their associated biological processes and molecular functions (based on Gene Ontology terms). We identified a set of processes or functions that are common to all these cancers, as well as those that are specific to only one or partial cancer types. Similarly, protein interaction networks in nucleic acid metabolism were compared to identify the common/specific clusters of proteins across different cancer types. Our results provide a basis for further experimental investigations to study protein interaction networks associated with cancer. The methodology developed in this work can also be applied to study similar disease systems.

Key words: protein-protein interaction, cancer-associated genes, GO similarity analysis, cancer bioinformatics

Introduction

The vast majority of proteins must interact with other proteins to perform their function. The network of protein-protein interactions (PPIs), referred to as the interactome, plays a vital role in the initiation and progression of many disease pathways. Hence, understanding protein interaction networks is crucial for identifying the key functional modulators of tumor progression and metastasis in cancer. Computational approaches using PPI data have been widely employed for the identification of important protein networks involved in tumors (*1*) and cancer metastasis (*2*). For instance, estrogen-regulated networks in human breast cancer cells were recently identified (*3*). Similarly, an earlier study on the interactome-transcriptome analysis has revealed the high centrality of differentially expressed genes in lung cancer tissues (*4*). Studies were also performed on integrated analysis of the cancer transcriptome using microarray data coupled with computational approaches (*5*). A recent study introduced a systems biology approach to improve the prognosis prediction of breast, lung and ovarian cancer patients (*6*), where the authors analyzed molecular interactions that are dysregulated in specific tumor phenotypes using a large set of microarray expression data. Likewise, protein interaction networks of brain metastasis were analyzed by developing and using a bioinformatics program called PIANA (*7*). These studies show the importance of computational methods in deducing and understanding protein interaction networks from datasets associated with cancer gene expression.

Analysis of microarray expression data to infer protein interaction networks has been the subject of extensive research aimed at gaining insight into the initiation and progression of various diseases like cancer. Several studies have been carried out on similar

lines using gene expression data from tumor tissues of ovarian cancer (*8*), prostate cancer (*9*), etc. Nevertheless, these studies were primarily focused on correlating gene expression data with PPIs occurring in a specific type of cancer. While these studies provide valuable information on the significance of a set of PPIs in a particular type of cancer, comparing and contrasting these interactions across various cancers will elucidate comprehensive knowledge on the nature of these interactions.

In the current study, we carried out global computational analysis of protein interactions in cancer-associated genes, using microarray expression datasets from normal and tumor tissues of four distinct cancer types, including bladder, colon, kidney and thyroid cancers. To the best of our knowledge, such kind of global computational analysis to compare PPIs across multiple cancer types has not been conducted before. The goals of this study are twofold: firstly, to identify and characterize PPIs that are common to all these cancers as well as those that are specific to a particular cancer type, and secondly, to infer the functional significance of PPIs by creating protein interaction networks using Gene Ontology (GO) annotations.

Results and Discussion

We implemented a top-down approach to carry out global analysis of the PPIs that are potentially involved in the initiation and progression of cancer. Figure 1 depicts the workflow of our methodology. We searched the GEO database (Gene Expression Omnibus) to collect datasets from gene expression studies where a tumor tissue (belonging to bladder, colon, kidney or thyroid cancers) was compared against a healthy tissue reference (see Materials and Methods for more details). To maintain consistency, the data for all cancer types analyzed in this study were collected from experiments that used only Affymetrix HG-U133 Plus 2.0 GeneChip. Due to a paucity of tumor-grade-specific datasets on this new Affymetrix GeneChip platform for all cancer types, we restricted our analysis to the baseline or essential PPIs in different cancer tumors (irrespective of the tumor grade) compared with normal tissues.

As a result, we identified a total of 6,758 unique proteins in the four cancer types, corresponding to the genes that are differentially expressed by at least two folds. Differentially expressed genes were first mapped to corresponding proteins and then to proteins involved in known PPIs. About 54% (3,656) of these proteins were mapped to 23,619 unique PPIs and were used for further analysis in this study. Mapped PPIs for each cancer were annotated using GO terms that describe their biological processes (pathways) and molecular functions. We compared PPIs from different cancer types using the frequency of common GO terms between two partner proteins in an interaction. The rationale is that the common GO terms of two partner proteins in an interaction denote the common biological process and/or molecular function associated with that interaction. Hence, based on the most frequent common GO terms in the PPIs of a particular cancer type, the nature of biological processes and/or functions associated with that cancer can be inferred. However, the incomplete and unbalanced nature of the available PPI datasets resulted in unbalanced datasets of mapped PPIs across the four cancers used in this study (Table 1). To make a fair comparison of GO term frequencies, we normalized the raw frequencies against the number of PPIs, showing at least one common GO term under a given GO category in a given cancer type. Accordingly, the top 20 most frequent common GO terms were collected from each cancer type and then were combined to obtain 30 unique GO terms from all these cancers, under the "biological process" and the "molecular function" categories, respectively. These GO terms were then used for comparative analysis of biological processes and molecular functions associated with the PPIs of different cancer types.

Table 1 Statistics on the size of datasets across four different cancer types

Cancer	Total proteins mapped from DEGs*	Unique proteins with known interactions	Unique proteins having PPIs*	Unique PPIs*
Bladder	3,704	3,008	2,081	15,522
Colon	3,515	2,922	1,944	13,099
Kidney	1,257	1,034	728	4,978
Thyroid	1,380	1,124	715	4,599

*DEGs, differentially expressed genes; PPIs, protein-protein interactions.

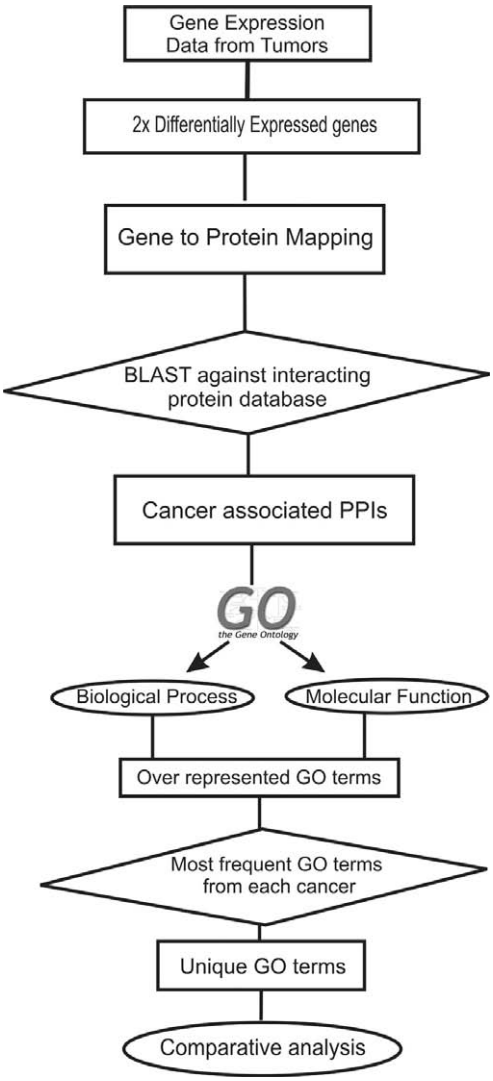

Figure 1 Flow chart showing the methodology used in this study.

Comparative analysis of biological processes associated with cancer PPIs

The most frequent biological processes associated with differentially expressed genes in the four cancers are listed in Table 2. As shown in Figure 2A, there are 14 common regulatory processes found in the differentially expressed genes of all these cancers. We further analyzed the types of PPIs associated with only one or two types of cancers (Figure 2B). Note that a GO biological process observed in only a particular cancer does not necessarily mean that such process is entirely lacking in other cancer types, rather it is not found in the list of experimentally known PPI datasets, which are incomplete and unbalanced for various types of cancers. Our methodology found a number of interesting biological processes that are specific to a particular cancer type. For instance, cell-cell signaling (CCS) and macromolecule catabolic process (MCP) were observed only in bladder cancer, which are linked to signaling proteins like hormones and glycoproteins that are mostly oncogenes (*10*). The TEK receptor tyrosine kinase, which is associated with angiogenesis (a primary process that establishes blood supply for the tumors), is one of the proteins in cell signaling process (*11*). Inhibin is another signaling protein whose levels have been shown to reflect the size of cell tumors (*12*). However, experimental evidence suggesting a direct link between these proteins and bladder cancer is not available. Protein interactions associated with blood coagulation (BC) and hemostasis (HE) were observed only in thyroid cancer. Literature evidence shows that disturbance of hemostasis is a common phenomenon in patients with thyroid disease, usually with hyperthyroidism (*13*). In these studies, a hyper coagulability state has been

Table 2 Gene Ontology terms (biological process) of differentially expressed genes

Processes common to all of the four cancer types		
ASM	anatomical structure morphogenesis	
BMP	biopolymer metabolic process	
CD	cell differentiation	
CL	cellular localization	
CMM	cellular macromolecule metabolic process	
NNN	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	
NRBP	negative regulation of biological process	
OD	organ development	
OOB	organelle organization and biogenesis	
PP	phosphorus metabolic process	
PRB	positive regulation of biological process	
RCM	regulation of cellular metabolic process	
SD	system development	
ST	signal transduction	
Processes found only in specific cancers		
BC	blood coagulation	thyroid
CBP	cellular biosynthetic process	colon, thyroid
CCS	cell-cell signaling	bladder
CP	cell cycle process	bladder, kidney, thyroid
GPM	generation of precursor metabolites and energy	colon
HE	hemostasis	thyroid
HOP	homeostatic process	colon, kidney
MCP	macromolecule catabolic process	bladder
MOB	membrane organization and biogenesis	kidney
PMP	protein metabolic process	colon, thyroid
PT	protein transport	colon
RMP	regulation of membrane potential	kidney
RST	regulation of signal transduction	bladder, kidney
RW	response to wounding	bladder, thyroid
SEM	symbiosis, encompassing mutualism through parasitism	colon
VMT	vesicle-mediated transport	bladder, kidney

reported in patients with thyroid cancer as a result of up-regulation of thrombin, protein S alpha and coagulation factor X. Similarly, PPIs involved in membrane organization and biogenesis (MOB) and regulation of membrane potential (RMP) were observed only in kidney cancer. Proteins associated with these processes regulate ionic homeostasis, suggesting the importance of homeostasis in renal cell carcinoma (*14*). Moreover, most of these interactions are among membrane proteins that are vital for membrane transport to maintain homeostasis (chloride channels, Na^+/H^+ exchanger, ferritin, the iron storing protein, Thy-1 membrane glycoprotein, etc).

Lastly, PPIs associated with the generation of precursor metabolites and energy (GPM), symbiosis encompassing mutualism through parasitism (SEM) and protein transport (PT) were observed only in colon cancer. Proteins associated with the GPM process are mostly mitochondrial respiratory enzymes and mitochondrial outer membrane proteins. It is reported that a quantitative mitochondrial change occurs in colon carcinoma patients; in particular, the ratio between outer membrane enzyme activity and respiratory enzyme activity may be altered (*15*, *16*). Another process that is regulated only in colon cancer is SEM. This is not surprising because the role of bacteria in the development of colorectal tumors was well documented (*17*). The involvement of colonic bacteria expressing enzymes in the metabolism of procarcinogens and tumor promoters suggests that changes

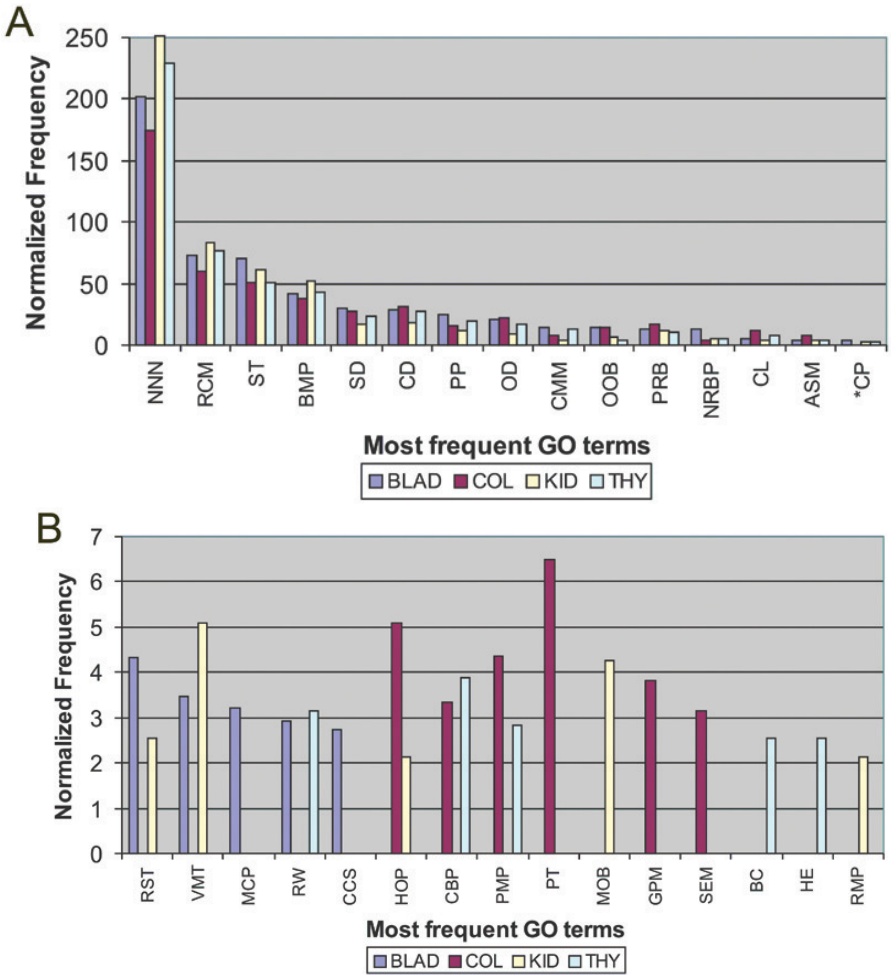

Figure 2 Comparison of the frequency distribution of GO terms describing the common biological processes in the PPIs of different cancers including bladder (BLAD), colon (COL), kidney (KID) and thyroid (THY) cancers. **A**. Histogram showing the most frequent processes in at least three types of cancers; processes found only in three types of cancer are marked with asterisk. **B**. Histogram showing the most frequent processes in only one or two types of cancers.

in the colonic bacterial population would influence cancer risk by altering the activity of these enzymes (*18*). Major proteins involved in the SEM process are ubiquitin protein ligase and HLA class histocompatibility antigen. Proteins associated with the PT process are chromatin modifying protein, SH3 and PX domain containing protein (sorting nexin-9), which are involved in intracellular trafficking. The chromatin remodeling enzymes with their role in cancer and dysplastic syndromes were previously reported (*19*). Studies on sorting of nexin-9 and EG-1 proteins that are significantly elevated in colorectal, breast and prostate cancer revealed the possible role of nexin-9 in signaling pathway (*20*).

Other processes regulated in multiple cancers include regulation of signal transduction (RST) and vesicle-mediated transport (VMT) in bladder and kidney cancers, homeostatic process (HOP) in colon and kidney cancers, regulation of protein metabolic process (PMP) and cellular biosynthetic process (CBP) in colon and thyroid cancers, and response to wounding (RW) in bladder and thyroid cancers. PPIs associated with the RW process were observed among proteins such as thrombin, serine proteinase inhibitor, prostaglandin endoperoxidase synthase (aka COX-2, cyclooxygenase-2), which are all responsible for inducing tumor growth, metastasis and angiogenesis (*21*, *22*). Nevertheless, we were unable to find literature-based explanations for each and every observation in our study, yet the aforementioned experimental evidence strongly suggests that the methodology used in this study is able to identify the true biological processes involved in the PPIs of cancer-associated genes.

Comparative analysis of molecular functions associated with cancer PPIs

We also carried out similar analysis of GO terms that describe the common "molecular functions" of PPIs associated with different cancers. We collected the top 20 most frequent GO terms from the PPIs of each of the four cancers and then combined them to obtain 30 unique GO terms describing their molecular functions. Out of the 30 terms, 12 ones are common to all these cancers (Figure 3A), including functions related to the binding of protein, ion, carbohydrate and nucleic acid, as well as to the catalytic activity (listed in Table 3). This result clearly indicates that binding and catalytic activities are the most common functions associated with all types of cancers. Meanwhile, some functions were found in the PPIs of only three cancer types. For instance, two functions, hydrolase activity acting on ester bonds (HAE) and protein binding, bridging (PBB), were regulated in all these cancers except in kidney cancer; PPIs associated with alpha-type channel activity (ATC), hydrolase activity acting on acid anhydrides (HAA) and identical protein binding (IPB) were expressed in all these cancers except in thyroid cancer; while protein domain specific binding (PDSB) function was not found in the PPIs of bladder cancer.

Similar to the biological processes, we have identified several molecular functions that are specifically associated with a particular type of cancer (Figure 3B). Proteins regulated only in bladder cancer include kinases and synthetases that belong to the GO cate-

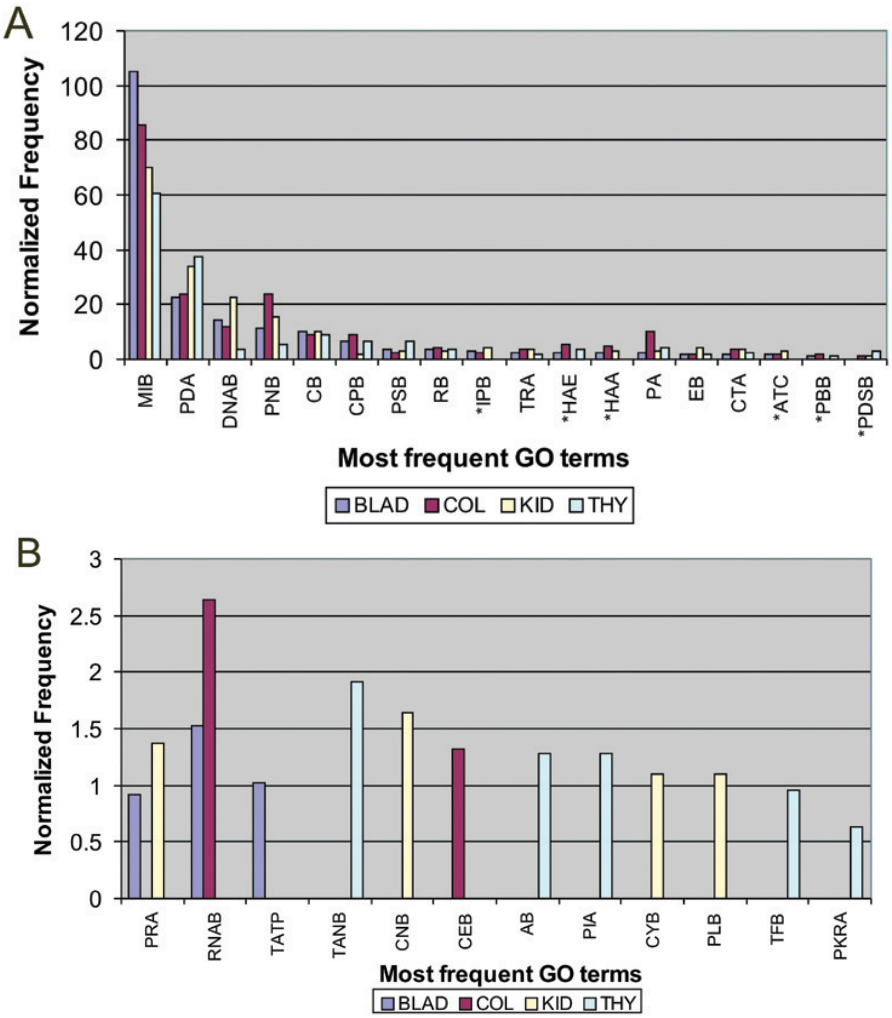

Figure 3 Comparison of the frequency distribution of GO terms describing the common molecular functions in the PPIs of different cancers including bladder (BLAD), colon (COL), kidney (KID) and thyroid (THY) cancers. **A**. Histogram showing the most frequent functions in at least three types of cancers; functions found only in three types of cancer are marked with asterisk. **B**. Histogram showing the most frequent functions in only one or two types of cancers.

Table 3 Gene Ontology terms (molecular function) of differentially expressed genes

Functions common to all of the four cancer types		
CB	cation binding	
CPB	cytoskeletal protein binding	
CTA	cation transport activity	
DNAB	DNA binding	
EB	enzyme binding	
MIB	metal ion binding	
PA	peptidase activity	
PDA	protein dimerization activity	
PNB	purine nucleotide binding	
PSB	polysaccharide binding	
RB	receptor binding	
TRA	transmembrane receptor activity	
Functions found only in specific cancers		
AB	anion binding	thyroid
ATC	alpha-type channel activity	bladder, colon, kidney
CEB	coenzyme binding	colon
CNB	cyclic nucleotide binding	kidney
CYB	cytokine binding	kidney
HAA	hydrolase activity, acting on acid anhydrides	bladder, colon, kidney
HAE	hydrolase activity, acting on ester bonds	bladder, colon, thyroid
IPB	identical protein binding	bladder, colon, kidney
PBB	protein binding, bridging	bladder, colon, thyroid
PDSB	protein domain specific binding	colon, kidney, thyroid
PIA	protease inhibitor activity	thyroid
PKRA	protein kinase regulator activity	thyroid
PLB	phospholipids binding	kidney
PRA	peptide receptor activity	bladder, kidney
RNAB	RNA binding	bladder, colon
TANB	translation factor activity, nucleic acid binding	thyroid
TATP	transferase activity, transferring phosphorus-containing groups	bladder
TFB	transcription factor binding	thyroid

gory of transferase activity, transferring phosphorus containing groups (TATP). There is concrete evidence on the role of these enzymes in tumorigenesis and metastasis (*23*, *24*). Likewise, cyclic nucleotide binding (CNB), cytokine binding (CYB) and phospholipid binding (PLB) functions are involved only in the PPIs observed from renal cell carcinoma. Proteins regulated in this cancer include chemokine (a type of cytokine) receptors and contactin-1 protein (a phospholipid binding protein). A relationship between cytokine binding and renal cancer cells was reported in a previous study (*25*), and the role of contactin-1 was found to be essential in tumor invasion and metastasis (*26*). PPIs involved in coenzyme binding (CEB) function were regulated only in colon cancer. Proteins with this function are mostly mitochondrial, and the relationship between mitochondrial enzymes and colon cancer was previously explained in the biological process section (*15*). Molecular functions including protease inhibitor activity (PIA), anion binding (AB), transcription factor binding (TFB), translation factor activity, nucleic acid binding (TANB) and protein kinase regulatory activity (PKRA) were found to regulate only in the PPIs of thyroid cancer. Proteins associated with these functions include cathepsin, runt-related transcription factor, protein kinase, cycline-dependent kinase inhibitor, etc. The regulation of these proteins and their role in tumorigenesis was well documented (*27*–*29*).

We also observed some molecular functions that

are seen in only two types of cancers. For instance, peptide receptor activity (PRA) was found only in bladder and kidney cancers. Proteins involved in these interactions include signaling receptors like G-protein coupled receptor, interleukin-8 receptor, chemokine and bradykinin receptors. Expression and regulation of these proteins was reported in bladder cancer (*30*, *31*).

Analysis of protein interaction networks in nucleic acid metabolism

To demonstrate the effectiveness of this method in identifying networks of genes that have different regulatory roles in different cellular processes associated with cancer, we chose the biological process NNN (nucleobase, nucleoside, nucleotide and nucleic acid metabolism) for analysis, which is highly up-regulated in all the four cancer types used in this study. We created protein interaction networks for each cancer separately, using PPIs that are associated with the NNN process. These networks were compared against each other to identify the similarities and differences among them. As shown in Figure 4, we found that some networks are universal to all these cancers (Figure 4A), while some are found only in part of these cancers (Figure 4B) or only in a specific type of cancer such as bladder cancer (Figure 4C) or colon cancer (Figure 4D). The observations are elucidated as follows.

Oncogenic activation of transcription factors is a key event in the establishment and progression of human cancers. Most of the members of the networks in Figure 4 are either transcription factors or their activators/receptors. Figure 4A shows a network of transcription factor activating proteins that are common to all of the four cancer types. The activity of these proteins was shown to increase in multiple human tumor types, suggesting their pivotal role in tumorigenesis (*32*). These activator proteins were recognized as molecular targets for many antioxidant and anti-inflammatory chemopreventive compounds. For example, the transcription factor activation protein 1 (JUN) is associated with progression and recurrence of various types of cancers (*33*), and several nuclear receptors (NR1P1, NCOA2, NR4A1) play a pivotal role in controlling the growth and differentiation in many cell types. These proteins are also known to inhibit or enhance transcription by recruiting an array of co-activator or co-repressor proteins to the transcription complex (*34*). Transcription factors are generally not suitable as drug targets except the nuclear receptors, which are considered as outstanding targets for developing cancer therapeutics and drug discovery (*35*). Other proteins in this network include hypoxia-inducible factor 1 alpha (HIF1A), which, in association with other proteins, may be involved in angiogenesis and tumor growth (*36*).

Figure 4B shows the proteins containing WW domain (WWTR1) or homeobox domain (NKX2.1) that

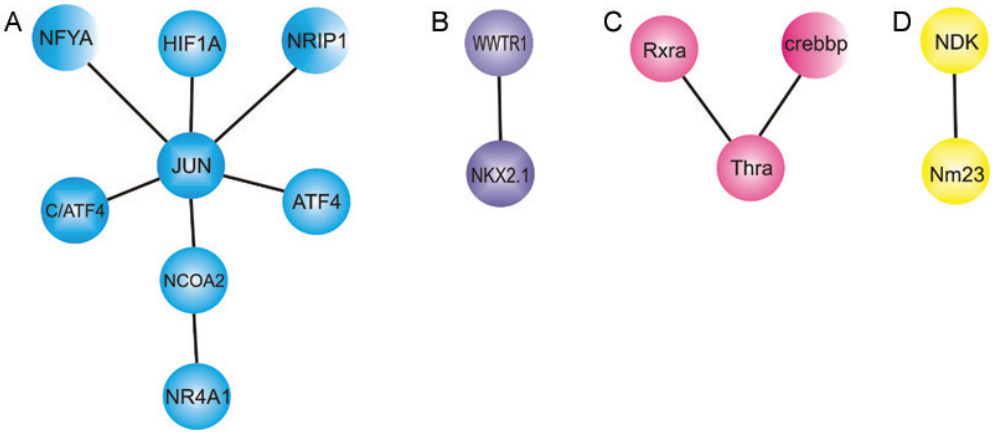

Figure 4 Protein interaction networks of nucleic acid pathway. **A**. NFYA, nuclear transcription factor Y subunit alpha; HIF1A, hypoxia inducible factor 1 alpha; NRIP1, nuclear receptor interacting protein 1; JUN, transcription factor activator protein 1; NCOA2, nuclear receptor co-activator 2; NR4A1, nuclear receptor subfamily 4 group A member 1; ATF4, activating transcription factor 4; C/ATF4, cyclic AMP-dependent transcription factor ATF-4. **B**. WWTR1 (TAZ), WW domain-containing transcription regulator protein; NKX2.1, homeobox protein Nkx-2.1. **C**. Rxra, retinoic acid receptor RXR-alpha; crebbp, CREB binding protein; Thra, thyroid hormone receptor alpha. **D**. NDK, nucleoside diphosphate kinase; Nm23, nucleoside diphosphate kinase, mitochondrial (precursor).

are specific to both colon and bladder cancers. Though their expression and involvement are not well studied in bladder and colon cancers, their critical role was reported in breast cancer (*37*) (not included in this study due to constraints on HG-U133 plus 2 GeneChip data availability). The network that is specific to bladder cancer (Figure 4C) contains important receptor proteins, whose role as biomarkers was well studied for developing therapeutic drugs in bladder cancer prevention (*38*, *39*). The network in Figure 4D was found only in colon cancer, which contains two nucleoside diphosphate kinases NDK and Nm23. Out of them, the mitochondrial precursor protein (Nm23) was well studied as a metastasis-associated gene in colon cancer (*40*).

We also looked into the subcellular localization of the proteins involved in these networks. The result shows that almost all of these proteins are localized in the nucleus except those in Figure 4D, where Nm23 is localized to mitochondria and the localization of NDK is unknown. The nuclear localization of these proteins is expected owing to their role in the regulation of transcription. In addition, comparison of PPI networks across four cancers and identification of experimentally known protein clusters that are common or specific to different cancer types have demonstrated the efficacy of our method in studying similar interaction networks in other disease systems.

Conclusion

We performed bioinformatic analysis of differentially expressed genes obtained from microarray studies of various tumor tissues. Comparative analysis of PPIs from different cancer types revealed a number of common functions or processes across all these cancers, as well as those that are specific to partial cancers or only to a particular cancer type. The function of a protein in a system (such as a tumor) is better understood by studying the function(s) of its interacting proteins. The methodology used in this study derives the common functions of protein pairs in PPIs from different tumor tissues and uses this formation as the basis for cross comparison of similarities and differences among various cancer types. Comparison of protein interaction networks revealed the group of genes that are regulated uniformly across all these cancer types as well as those regulated only in a specific cancer, indicating their importance in that particular cancer. We have provided literature-based evidence for many observations in this study, yet such evidence is not available to support every observation. We believe that the similarities and differences observed in the biological processes and molecular functions of PPIs from various cancer types will provide the basis for focused experimental investigations in cancer therapeutics and drug discovery studies.

Materials and Methods

Cancer gene expression data

The GEO database (*41*), a central repository for differential gene expression data, was used in this study. This database offers an extensive collection of gene expression data on cancer and compares various types and subgroups. We selected only those gene expression studies where a tumor tissue was compared against a healthy tissue reference. The initial datasets were derived for six cancer types including bladder (GSE7476), colon (GSE4107), kidney (GSE7023), thyroid (GSE3678), breast (GSE6883) and prostate (GSE3325) cancers. To maintain consistency, data for all cancer types in this study were collected from experiments that used only Affymetrix HG-U133 Plus 2.0 GeneChip. This is a newer chip from Affymetrix that covers about 54,000 human genes, but the tradeoff is that this restriction limits the number of available datasets on a specific tumor grade for a specific cancer. Since the goal of our study is to carry out comparison of baseline PPIs in different cancer types, it requires robust datasets for each cancer type. To obtain datasets of reasonable size, we didn't take the tumor grade into account; hence heterogeneity somewhat exists in the grade of the tumor tissue samples used in this study. In brief, bladder, kidney and prostate datasets were obtained from low, high or metastatic (invasive) tumors, whereas breast, colon and thyroid samples originated mostly from low grade or early onset tumors. Under this GeneChip, breast cancer and prostate cancer were found to have only 34 and 139 differentially expressed genes, respectively; hence these two cancer types were eliminated from further analysis. More details on the grade and tumor type used for each cancer are provided in Table S1 (see Supporting Online Material).

Datasets in each series were analyzed by GeneSpring software (v7.3.1; Agilent) using GCRMA normalization and with the cross-gene error model activated. In each case, samples were grouped as normal or diseased, and a parametric t-test was per-

formed to obtain genes with statistically significant differences. The p-value cutoff was set at 0.05 and a Benjamin Hochberg False Discovery Rate correction was included. These restrictions would allow about 5% of the tested genes on the GeneChip (total 54,675) to qualify by random chance. This gene list was further filtered to obtain genes that showed at least a two-fold differential expression between the normal and diseased states, which generated between 1,300 to 3,700 differentially expressed genes for the four cancer datasets (Table 1). Protein sequences corresponding to these genes were obtained from GenBank and UniProt databases.

Protein-protein interaction data

We created a comprehensive, non-redundant dataset of human interacting protein pairs (PPIs) by combining experimentally derived datasets from five major protein interaction databases, including DIP (Database of Interacting Proteins; http://dip.doe-mbi.ucla.edu), IntAct (http://www.ebi.ac.uk/intact), BIND (Biomolecular Interaction Network Database; http://www.bind.ca), HPRD (Human Protein Reference Database; http://www.hprd.org) and MINT (Molecular Interaction Database; http://mint.bio.uniroma2.it/mint). Protein sequences in each of these databases are fairly overlapping, but are indexed by different identifiers such as SwissProt/UniProt identifier, GenBank protein identifier, etc, making it difficult to identify redundant sequences. To remove redundancy, we first created datasets of unique sequences (based on full-length protein sequence string comparison using Perl scripts) within each database and then merged them to create a non-redundant dataset of interacting protein sequences, each indexed with our internal identifier. Note that this internal identifier can be used to map all the original source identifiers for corresponding sequences in the source databases. Finally, we obtained 27,051 unique human protein sequences (henceforth referred to as P_{int}) representing 57,307 unique PPIs. The FASTA formatted protein sequences and PPI datasets for each cancer used in this study can be downloaded at http://bioinformatics.albany.edu/gpb/.

Mapping and annotation of interacting proteins

To map the interacting proteins with corresponding differentially expressed genes, we performed a BLAST search against the P_{int} dataset described above using very stringent criteria as following. For each query protein, the top hits with a sequence identity of $\geq 95\%$ and a sequence length match between 90%–110% were selected. Functional annotations for the selected interacting proteins were obtained from the GO database (http://www.geneontology.org). We developed several Perl scripts to analyze the most frequent common GO terms in PPIs belonging to individual cancer types.

Comparative analysis of PPIs based on GO term frequency

GO is a hierarchical graph-based annotation system where the terms closer to the root describe more general information while those away from the root provide more specific information about a given GO category. The root (level 0) describes three main GO categories at level 1, which are "biological process", "molecular function" and "cellular component". We used GO terms only from the "biological process" and "molecular function" categories because the data for the "cellular component" category are very sparse. Ideally, we would have used GO terms that provide specific descriptions on the processes or functions; however, the more specific the terms get, the less frequent they are, which prohibits meaningful comparison of GO term frequencies across different cancer. Hence, we chose the GO terms at level 4 for our analysis because at this level, GO terms are more specific (than those at previous levels), while generic enough to cover broader groups of related processes or functions to attain reasonable cumulative frequency for analysis. All the GO terms associated with a protein sequence were obtained from the GO database as graph paths, which have an inherent hierarchical order starting from the root (level 0). Since our goal is to compare the number of common GO terms across different cancer types, these GO terms must be selected at the same level in the graph path (for all cancers) to make the comparison meaningful. We developed a Perl program to store all the graph paths for a given protein. Using this program, under each GO category (except the "cellular component"), the common GO terms for a given pair of PPIs were determined at level 4 to calculate the frequency of common GO terms in a given cancer. Since the size of PPI dataset (Table 1) varies across different cancer types, frequencies of common GO terms were normalized against the number of PPIs, showing at least one

common GO term in each cancer dataset. This normalization ensures a fair comparison of the common GO term frequencies across different cancer types irrespective of the size of PPI datasets used. The top 20 most frequent common terms were selected from each cancer, which were combined to obtain a non-redundant set of GO terms under the biological process and the molecular function categories, respectively.

Creation of protein interaction networks

PPIs associated with differentially expressed cancer genes were used to create networks with the application of Cytoscape program (v2.3.2) (http://www.cytoscape.org). Semantic similarity provides a quantitative measure of how similar a pair of proteins are, based on the annotations (GO terms) in a given GO category. This method has been proved to be very effective in interpreting the functional similarities of genes based on gene annotation information from heterogeneous data sources (*42*, *43*). In our study, the semantic similarity between the molecular functions or the biological processes of two proteins involved in a PPI was calculated following the literature (*43*) and only those PPIs with a high semantic similarity value of 4.0 or more (5.5 being the highest) were used to create interaction networks under each GO term category.

Acknowledgements

This work was partly supported by the start-up funds to CG from SUNY-Albany, and partly by the Academic Research Enhancement Award (1R15GM080681-01) to CG from NIGMS/NIH.

Authors' contributions

PG carried out this work, analyzed the results and drafted the manuscript. SVC created the microarray datasets and contributed to the manuscript preparation and proofreading. CG conceived of the study, provided overall conceptual framework for this project and contributed to the manuscript preparation. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Platzer, A., *et al.* 2007. Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* 8: 224.
2. Jonsson, P.F., *et al.* 2006. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* 7: 2.
3. Stender, J.D., *et al.* 2007. Estrogen-regulated gene networks in human breast cancer cells: involvement of E2F1 in the regulation of cell proliferation. *Mol. Endocrinol.* 21: 2112-2123.
4. Wachi, S., *et al.* 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21: 4205-4208.
5. Rhodes, D.R. and Chinnaiyan, A.M. 2005. Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37: S31-37.
6. Mani, K.M., *et al.* 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol. Syst. Biol.* 4: 169.
7. Martín, B., *et al.* 2008. Biological pathways contributing to organ-specific phenotype of brain metastatic cells. *J. Proteome Res.* 7: 908-920.
8. Chen, J.Y., *et al.* 2006. A systems biology case study of ovarian cancer drug resistance. *Comput. Syst. Bioinformatics Conf.* 389-398.
9. Ergün, A., *et al.* 2007. A network biology approach to prostate cancer. *Mol. Syst. Biol.* 3: 82.
10. Stein, J.P., *et al.* 1998. Prognostic markers in bladder cancer: a contemporary review of the literature. *J. Urol.* 160: 645-659.
11. Sun, L. and McMahon, G. 2000. Inhibition of tumor angiogenesis by synthetic receptor tyrosine kinase inhibitors. *Drug Discov. Today* 5: 344-353.
12. Lappöhn, R.E., *et al.* 1989. Inhibin as a marker for granulosa-cell tumors. *N. Engl. J. Med.* 321: 790-793.
13. Franchini, M. 2006. Hemostatic changes in thyroid diseases: haemostasis and thrombosis. *Hematology* 11: 203-208.
14. Boer, J.M., *et al.* 2001. Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res.* 11: 1861-1870.

15. Sun, A.S., *et al.* 1981. A study of some mitochondrial and peroxisomal enzymes in human colonic adenocarcinoma. *Lab. Invest.* 44: 13-17.
16. Washo-Stultz, D., *et al.* 2002. Role of mitochondrial complexes I and II, reactive oxygen species and arachidonic acid metabolism in deoxycholate-induced apoptosis. *Cancer Lett.* 177: 129-144.
17. Yang, L. and Pei, Z. 2006. Bacteria, inflammation, and colon cancer. *World J. Gastroenterol.* 12: 6741-6746.
18. Gallaher, D.D. and Khil, J. 1999. The effect of synbiotics on colon carcinogenesis in rats. *J. Nutr.* 129: 1483S-7S.
19. Gibbons, R.J. 2005. Histone modifying and chromatin remodelling enzymes in cancer and dysplastic syndromes. *Hum. Mol. Genet.* 14: R85-92.
20. Lu, M., *et al.* 2006. EG-1 interacts with c-Src and activates its signaling pathway. *Int. J. Oncol.* 29: 1013-1018.
21. Nierodzik, M.L. and Karpatkin, S. 2006. Thrombin induces tumor growth, metastasis, and angiogenesis: evidence for a thrombin-regulated dormant tumor phenotype. *Cancer Cell* 10: 355-362.
22. Badawi, A.F., *et al.* 2002. Influence of cigarette smoking on prostaglandin synthesis and cyclooxygenase-2 gene expression in human urinary bladder cancer. *Cancer Invest.* 20: 651-656.
23. Franzon, V.L, *et al.* 1999. Molecular cloning of a novel human PAPS synthetase which is differentially expressed in metastatic and non-metastatic colon carcinoma cells. *Int. J. Biochem. Cell Biol.* 31: 613-626.
24. Bhola, N.E. and Grandis, J.R. 2008. Crosstalk between G-protein-coupled receptors and epidermal growth factor receptor in cancer. *Front. Biosci.* 13: 1857-1865.
25. Steinbach, F., *et al.* 1996. The influence of cytokines on the adhesion of renal cancer cells to endothelium. *J. Urol.* 155: 743-748.
26. Su, J.L., *et al.* 2006. Knockdown of contactin-1 expression suppresses invasion and metastasis of lung adenocarcinoma. *Cancer Res.* 66: 2553-2561.
27. Srisomsap, C., *et al.* 2002. Detection of cathepsin B up-regulation in neoplastic thyroid tissues by proteomic analysis. *Proteomics* 2: 706-712.
28. Asou, N. 2003. The role of a Runt domain transcription factor AML1/RUNX1 in leukemogenesis and its clinical implications. *Crit. Rev. Oncol. Hematol.* 45: 129-150.
29. Kogai, T., *et al.* 2006. Enhancement of sodium/iodide symporter expression in thyroid and breast cancer. *Endocr. Relat. Cancer* 13: 797-826.
30. Inoue, K., *et al.* 2000. Interleukin 8 expression regulates tumorigenicity and metastasis in human bladder cancer. *Cancer Res.* 15: 2290-2299.
31. Eisenhardt, A., *et al.* 2005. Association study of the G-protein beta3 subunit C825T polymorphism with disease progression in patients with bladder cancer. *World J. Urol.* 23: 279-286.
32. Matthews, C.P., *et al.* 2007. AP-1 a target for cancer prevention. *Curr. Cancer Drug Targets* 7: 317-324.
33. Ouyang, X., *et al.* 2008. Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer. *Cancer Res.* 68: 2132-2144.
34. Horwitz, K.B., *et al.* 1996. Nuclear receptor coactivators and corepressors. *Mol. Endocrinol.* 10: 1167-1177.
35. Ao, A., *et al.* 2008. Involvement of estrogen-related receptors in transcriptional response to hypoxia and growth of solid tumors. *Proc. Natl. Acad. Sci. USA* 105: 7821-7826.
36. Nakayama, K., *et al.* 2002. Hypoxia-inducible factor 1 alpha (HIF-1 alpha) gene expression in human ovarian carcinoma. *Cancer Lett.* 176: 215-223.
37. Chan, S.W., *et al.* 2008. A role for TAZ in migration, invasion, and tumorigenesis of breast cancer cells. *Cancer Res.* 68: 2592-2598.
38. Boorjian, S., *et al.* 2005. Retinoid receptor mRNA expression profiles in human bladder cancer specimens. *Int. J. Oncol.* 26: 1041-1044.
39. Hemstreet, G.P. 3rd, *et al.* 2001. Biomarker risk assessment and bladder cancer detection in a cohort exposed to benzidine. *J. Natl. Cancer Inst.* 93: 427-436.
40. Mönig, S.P., *et al.* 2007. Clinical significance of nm23 gene expression in gastric cancer. *Anticancer Res.* 27: 3029-3033.
41. Barrett, T., *et al.* 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.* 35: D760-765.
42. Brown, K.R. and Jurisica, I. 2005. Online predicted human interaction database. *Bioinformatics* 21: 2076-2082.
43. Wang, J.Z., *et al.* 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23: 1274-1281.

Supporting Online Material

Table S1

DOI: 10.1016/S1672-0229(08)60030-3