A SYSTEMS BIOLOGY CASE STUDY OF OVARIAN CANCER DRUG RESISTANCE

Jake Y. Chen[1,2,*], Changyu Shen[3], Zhong Yan[1], Dawn P. G. Brown[4], Mu Wang[4]

[1] Indiana University School of Informatics, IUPUI, Indianapolis, IN 46202; [2] Department of Computer and Information Science, Purdue University School of Science, IUPUI, Indianapolis, IN 46202; [3] Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202; [3] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202 [*] Corresponding Author. Email: jakechen@iupui.edu

In ovarian cancer treatment, the chemotherapy drug cisplatin often induce drug resistance after prolonged use, causing cancer relapse and the eventual deaths of patients. Cisplatin-induced drug resistance is known to involve a complex set of cellular changes but its molecular mechanism(s) remain unclear. In this study, we designed a systems biology approach to examine global protein level and network level changes by comparing Proteomics profiles between cisplatin-resistant cell lines and cisplatin-sensitive cell lines. First, we used an experimental proteomics method based on a Label-free Liquid Chromatography / Mass Spectrometry (LC/MS) platform to obtain a list of 119 proteins that are differentially expressed in the samples. Second, we expanded these proteins into a cisplatin-resistant activated sub-network, which consists of 1230 proteins in 1111 protein interactions. An examination of network topology features reveals the activated responses in the network are closely coupled. Third, we examined sub-network proteins using Gene Ontology categories. We found significant enrichment of proton-transporting ATPase and ATP synthase complexes in addition to protein binding proteins. Fourth, we examined sub-network protein interaction function categories using 2-dimensional visualization matrixes. We found that significant cellular physiological responses arise from endogeneous, abiotic, and stress-related signals, which correlates well with known facts that internalized cisplatin cause DNA damage and induce cell stress. Fifth and finally, we developed a new visual representation structure for display of activated sub-networks using functional categories as network nodes and their crosstalk as network edges. This type of sub-network further shows that while cell communication and cell growth are generally important to tumor mechanisms, molecular regulation of cell differentiation and development caused by responses to genomic-wide stress seem to be more relevant to the acquisition of drug resistance.

1. INTRODUCTION

The study of cancer drug resistance entails great challenges and opportunities for cancer chemotherapy. In ovarian cancer, each year in the United States, approximately 23,000 women are diagnosed with the disease, and approximately 15,000 will die, making ovarian cancer to rank second only to breast cancer by the total number of new patients and first by the total number of deaths among all gynecological cancer cases each year. Platinum-based chemotherapy, usually with the cancer drug "cisplatin", has been the primary treatment for ovarian cancer today [1]. Cisplatin is known to bind DNA to form cisplatin-DNA adducts and therefore could inhibit DNA replication and/or transcription of cancer cells [2]. At the beginning of such chemotherapeutic treatments, patients are usually responsive. As the treatment goes on for six to twelve months, however, many patients would relapse into cancer and eventually die from spread of new tumors that are refractory to further cisplatin treatment even at large toxic doses [3]. In cisplatin-resistant cancer cells when compared with initial cisplatin-sensitive cancer cells, cancer researchers have observed a diverse range of cellular changes, which include decreased drug accumulation, increased cellular glutathione, and enhanced DNA repair capacity [3, 4]. Recent advances in microarray technology has enabled researchers to identify global differential gene expression patterns between cisplatin resistant and cisplatin sensitive cell lines [5]. Notable examples of differentially expressed genes are: DNA repair related genes (BRCA1, DNA-PK, and ERCC1), P-glycoproteins genes, genes encoding heat shock proteins (HSP27, HSP70), and copper transport protein encoding genes (CTR). Despite these progresses, the molecular mechanism(s) of acquired cisplatin resistance remain unclear. This difficulty reflects a general challenge in applying a single genomics or functional genomics technology platform to complex disease biology problems. For example, it is generally known that there is a low correlation (15-25%) between global gene expression level and protein expression levels in higher eukaryotes including cancer cells [6]. Very low correlation (merely above random level) between global gene co-expressions and protein protein interaction pairs has also been observed [7]. Such evidence suggests that the control mechanism for intracellular signaling networks may not be fully understood at the gene expression level. A holistic approach to collect and interpret global signal changes beyond the gene expression level would provide significant additional insights, which in turn could lead to the development of novel therapeutic strategies for cancer.

In this work, we adopt a systems biology approach for the study of ovarian cancer drug resistance molecular mechanisms. For systems biology, we refer to the simultaneous experimental measurement of global molecular expression data in perturbed cells and computational interpretation of them at the molecular interaction network level or above (adapted and expanded from [8]). We consider systems biology as a holistic study approach distinct from a pure Omics or bioinformatics-based method. We summarize its characteristics as the following. 1) Actual functional genomics or proteomics experiments should be done on specific biological conditions driven by specific biological hypothesis. 2) Experimental Omics data derived should be interpreted in existing genomics knowledge context based on integrated annotated genomics database and integrated bioinformatics data analysis methods. 3) High-level knowledge structures at the molecular network or pathway levels must be derived.

Our system biology study of cisplatin drug resistant ovarian cancer consists of the following three elements. The first element is experimental proteomics study, using a label free Liquid Chromatography/Mass Spectrometry (LC/MS)-based technology platform (see Methods) to identify differentially expressed proteins in replicated cisplatin-resistant vs. cisplatin-sensitive ovarian cell line samples. The second element is the annotation of experimental proteomics results using human genome annotation database (from Gene Ontology [9]), human protein interaction network database (from OPHID [10]), integrated statistical data analysis, network analysis, and information visualization methods. The third element is the representation of our discovery results with a network of activated biological processes to summarize underlying complex protein interaction networks. These elements interweave into a coherent functional picture for the first time of the global changes in ovarian cancer cisplatin-resistant cells.

Our work has the following significances. First, LC/MS-based proteomics results that compares cisplatin resistant with cisplatin sensitive ovarian cancer cells have not been previously performed/reported elsewhere. Many new proteins involved in the drug resistance were identified. Second, we described and used a set of novel systems biology informatics methods and, in particular, one that can identify “significantly interacting protein categories”, distinct from previous work of using GO annotations for gene classifications from Microarray results analysis [11]. Collectively these methods can be generalized to enable other similar systems biology studies, in which statistically significant experimental Omics results, public protein interactome data, and genome/proteome annotation database are integrated into an easy-to-interpret 2-dimensional visualization matrix. Third, we developed a unique molecular network visual representation scheme based on automatically data-mined significant biological process categories and significant between-category interactions. This representation scheme can enable bioinformatics scientists to summarize and infer essential relationships between networked proteins in many similar application domains beyond our case study in this work.

2. METHODS

2.1. Proteomics Methods

A2780 and 2008 cisplatin-sensitive human ovarian cancer cell lines and their resistant counterparts, A2780/CP and 2008/C13*5.25, were used in this study. Proteins were prepared and subjected to LC/MS/MS analysis as previously described [12]. There were two groups (two different parent cell lines), six samples per cell line, and two HPLC injections per sample. Samples were run on a Surveyor HPLC (ThermoFinnigan) with a C18 microbore column (Zorbax 300SB-C18, 1mm x 5cm). All tryptic peptides (100 μL, or 20 μg) were injected onto the column in random order. Peptides were eluted with a linear gradient from 5 to 45% acetonitrile developed over 120 min at a flow rate of 50 μL/min, eluant was introduced into a ThermoFinnigan LTQ mass spectrometer. The data were collected in the triple-play mode. The acquired data were filtered by proprietary software as described by Higgs et al.[12]. Database searching against IPI human database and NR-Homo Sapiens database was carried out using SEQUEST algorithm. Protein quantification was carried out using the LC/MS-based label-free proprietary protein quantification software licensed from Eli Lilly and Company [12]. Briefly, once the raw files are acquired from the LTQ, all total ion chromatogram (TIC) will be aligned by retention time. Each aligned peak should match parent ion, charge state, daughter ions (MS/MS data) and retention time (within 1 minute window). If any of these parameters were not matched, the peak will be disqualified from the quantification. The area under the curve (AUC) from individually aligned peak was measured, normalized, and compared for their relative abundance. All peak intensities are transformed to a $\log_2$ scale before quantile normalization [13]. If multiple peptides have the same protein identification then their quantile normalized $\log_2$ intensities are averaged to obtain $\log_2$ protein intensities. The $\log_2$ protein intensity is the final quantity that is fit by a separate ANOVA statistical model for each protein. $\log_2$ (Intensity) = overall mean + group effect (fixed) + sample effect (random) + replicate effect (random). Group effect refers to the effect caused by the experimental conditions or treatments being evaluated. Sample effect is the random effects from individual biological samples. It also includes the random effects from sample preparations. The replicate effect refers to the random effects from replicate injections of the same sample. All of the injections were in random order and the instrument was operated by the same operator. The inverse $\log_2$ of each sample mean was determined to resolve the fold change between samples. A summary of the overall process is shown as Fig. 1.

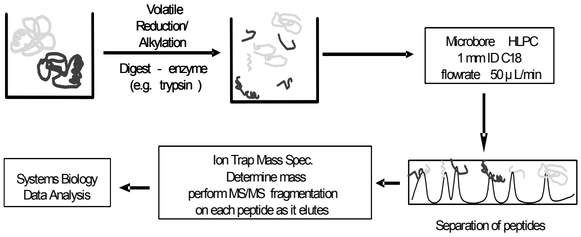

Figure 1. Schematic presentation of LC/MS based protein quantitative analysis of complex biological samples.

2.2. Preparation of Data Sets

Proteins in Differentially Expressed Cisplatin-Resistant vs. Cisplatin-Sensitive Ovarian Cancer Cells. Our experimental proteomics platform generated 574 differentially expressed proteins (with q-value <=0.10; both up- and down-regulation values) or 141 proteins (with q-value <=0.05), all identified with International Protein Index (IPI) database IDs. We convert these IPI identifiers into UniProt IDs in order to integrate this data set with all other annotated public data. 119 of the 141 proteins (0.05 q-value threshold) was successfully mapped and converted, using the International Protein Index (IPI) database [14] downloaded in March 2006, the UniProt database downloaded in November 2005 [15], and additional internally curated public database mapping tables. Similarly, 451 out of the 574 proteins with the less restrict threshold (q-value <=0.10) were mapped from IPI IDs to UniProt IDs.

Human Protein Interactome Data. The primary source of human data comes from the Online Predicted Human Interaction Database (OPHID) [10], which were downloaded in February 2006. It contains more than 47,213 human protein interactions among 10,577 proteins identified by UniProt accession numbers. After mapping the proteins in OPHID to UniProt IDs, we recorded 46,556 unique protein interactions among 9959 proteins. Note that even though more than half of OPHID are interacting protein pairs inferred from available lower organisms onto their human orthologous protein pair counterparts, the statistical significance of these predicted human interactions was confirmed by additional evidences according to OPHID and partially cross-validated according to our previous experience [16]. We assigned a heuristic interaction confidence score to each protein interactions, based on the types and sources proteins recorded in OPHID according to a method described in [16].

Human Protein Annotation Data. Gene Ontology (GO) classification data 9 were downloaded from Geneontology.org in January 2006 and were used as the primary source of protein annotation for this study. Human proteome GO annotation was further performed based on human gene GO annotation from NCBI and human gene ID to protein UniProt ID mappings curated internally.

Human Interacting Protein Categorical Annotation Data. Each GO term from the human protein annotation data was annotated with its minimal GO level number in the GO term hierarchy. Each GO term's higher-level parent GO terms (multiple parent GO terms are possible) up to GO level 1 (three GO terms at this level: molecular function, cellular components, and biological processes) are also traced and recorded in an internally curated GO annotation table. When calculating interacting protein GO category information, we use this internally curated GO term table to map all the low-level GO term IDs (original GO Term ID) used to annotate each protein to all the GO term IDs' high-level GO term IDs (folded GO Term ID). For this study, we designate that all the folded GO term ID should be at GO term hierarchy Level = 3. Note that our method allows for multiple GO annotation Term IDs (original or folded) generated for each protein ID on purpose. Therefore, it is possible for a protein or a protein interaction pair to appear in more than one folded GO term category or more than one folded GO term interacting category pairs.

2.3. Network Expansion

We derive ovarian cancer drug resistant-related differentially expressed protein interaction sub-network using a **nearest-neighbor expansion method** described in [16]. We call the original list of differentially expressed proteins (119 proteins) **seed (S) proteins** and all the protein interactions within them **seed interactions (**or **S-S type interactions)**. After expansion, we call the collection of seed proteins and expanded non-seed (N) proteins **sub-network proteins** (including both S and N proteins); we call the collection of seed interactions and expanded seed-to-non-seed interactions **(**or **S-N type interactions) sub-network protein interactions** (including both S-S type and S-N type interactions). Note that we do not include non-seed-to-non-seed protein interactions (or "**N-N**" type interactions) in our definition of the sub-network, primarily because the N-N type of protein interactions often outnumbered total S-S and S-N types of protein interaction by several folds with molecular network context often not tightly related to the initial seed proteins and seed interactions. The only occasion to consider the N-N type interactions is when we calculate sub-network properties such as node degrees for proteins in the sub-network.

2.4. Visualization

We use Spotfire DecisionSite Browser 7.2 to implement the 2-dimensional functional categorical crosstalk matrix. To perform interaction network visualization, we used ProteoLens [17]. ProteoLens has native built-in support for relational database access and manipulations. It allows expert users to browse database schemas and tables, query relational data using SQL, and customize data fields to be visualized as graphical annotations in the visualized network.

2.5. Network Statistical Examination

Since the seed proteins are those that are found to display different abundance level between two different cell lines via mass spectrometry, one would expect that the network "induced" by them to be more "connected" in the sense that they are to certain extent related to the same biological process(es). To gauge network "connectivity", we introduced several basic concepts. We define a **path** between two proteins A and B as a set of proteins P1, P2,…, Pn such that A interacts with P1, P1 interacts with P2, …, and Pn interacts with B. Note that if A directly interacts with B, then the path is the empty set. We define the **largest connected component** of a network, as the largest subset of proteins such that there is at least one path between any pair of proteins in the network. We define the **index of aggregation** of a network as the ratio of the size of the largest connected component of the network to the size of the network by protein counts. Therefore, the higher the index of aggregation, the more "connected" the network should be. Lastly, we define the **index of separation** of a sub-network as the ratio of S-S type interactions over all sub-network interactions. A high index of separation found in a network represents extensive "re-discovery" of proteins after the protein interactions are expanded from the seed proteins.

To examine the statistical significance of observed index of aggregation and index of separation in expanded protein networks, we measure the likelihood of the topology of the observed sub-network under random selection of seed proteins. This is done by randomly selecting 119 proteins, identifying the sub-network induced/expanded, and calculating sub-network indexes accordingly. The same procedure is repeated n=1000 times to generate the distribution of the indexes under random sampling, with which the observed values are compared to obtain significance levels (for details, refer to [18]).

2.6. Significance of Interacting Protein Categories

To assess the statistical significance on the number of pairs in the subnetwork that falls in specific function categories, we treat it as the outcome of a random draw of 1723 pairs from the pool of 46556 pairs in OPHID. Then it follows a hypergeometric distribution. A p-value is calculated based on the hypergeometric distribution to evaluate the likelihood that we observe an outcome under random selection of 1723 pairs that is at least as "extreme" as what we have observed. Note "extreme" either implies an unusual large (over-representation) or too small (under-representation) number. Let x be the count of the pair that falls in a function category in the subnetwork, n=1723, N=46556 and k=corresponding count in OPHID, then the p-value for over/under-representation of the observed count can be calculated as:

Over representation:

$$p = \Pr[X \ge x \mid n, N, k] = \sum_{i=x}^{\min(n,k)} \binom{k}{i}\binom{N-k}{n-i} \Big/ \binom{N}{n}$$

Under representation:

$$p = \Pr[X \le x \mid n, N, k] = \sum_{i=0}^{x} \binom{k}{i}\binom{N-k}{n-i} \Big/ \binom{N}{n}$$

Since tests of over/under representation of various categories are correlated with one another (over representation of one category could imply under representation of other categories), we also control the false discovery rate (FDR) using method developed by Benjamini and Yekutieli [19].

3. RESULTS

3.1. Activated Protein Interaction Sub-network Properties

We examined network topology for the protein interaction sub-network expanded from seed proteins. Recall that seed proteins are significantly and differentially expressed proteins derived from LC/MS proteomics experiments based on comparing cisplatin resistant with cisplatin sensitive cell line samples (see Methods). The resulting protein interaction sub-network consists of 1,230 seed and non-seed proteins in 1,723 sub-network interactions (including 17 S-S type interactions and 1,706 S-N type protein interactions). We call this protein interaction sub-network "core sub-network" to distinguish it from the "full sub-network" (additionally including all N-N type protein interactions), and plot their node degree frequency distributions in Fig 2, where the whole human protein interaction network from OPHID (labeled "network") is also shown. As expected, both the network and the sub-network (full) display good "scale-free" property. The result also shows cisplatin resistant activated sub-network (full) contains more "hubs" than "peripheral" proteins to form a cohesive functional sub-network. The core sub-network, while perhaps limited in size, begins to show "scale-free like" distribution, although hubs in the sub-network (core) are more distinctively identifiable than overly abundant peripheral nodes by high node degree counts.

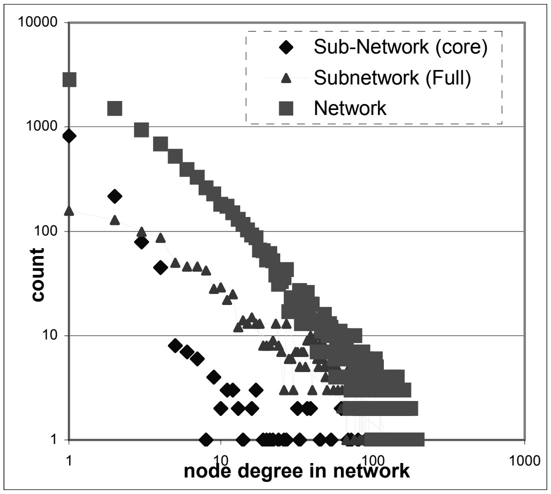

Figure 2. Node degree distribution of the sub-networks (core or full) in comparison with the human protein interaction network.

We also examined other network features for the core sub-network. The *largest connected component* (defined in the Method section; *ibid*) of the sub-network consists of 1230 proteins with 1723 interactions. The *index of aggregation* is 1193/1230=97.0%. The *index of separation* as the percentage of S-S type interactions (17) over the core sub-network interactions (1723), i.e., 17/1723=0.96%. The *index of aggregation* has a p-value less than 0.001 (upper tail) and the *index of separation* 0.06 (upper tail) A significant but not exceptionally high network *index of aggregation* suggests that the core sub-network has connectivity structures that are not random by nature. This correlates well with the previous node degree distribution in Fig. 2, where an exceptionally large number of hubs are shown to exist. A relative high (though not significant) *index of separation* after expansions suggests that the 119 seed proteins may be tightly related—an observation consistent with the assumption that majority of the connected proteins should participate in a few shared biological pathways defined by the activated sub-network.

3.2. Analysis of Activated Protein Functional Category Distributions

We are interested in discovering enriched protein functional categories among differentially expressed seed proteins and its immediate protein interaction sub-network nearest interaction partners. Note that this enrichment includes response ("activated") proteins that are either up-regulated or down-regulated in the proteomics experiment. Although this up-/down-regulation detail can be essential to establishing regulatory network models, we are more interested in proteins and protein groups that are "activated" in the cisplatin-response functional process than those that are "not activated". Therefore, we choose not to differentiate the regulation detail for this study.

Although GO-based functional category analysis can be done routinely using many existing bioinformatics methods [11], the inclusion of protein interaction network context has not been previously described. Here, we are interested in broad pathway-related changes in the sub-network, even this means the cost of failing to detect the significant appearance of some isolated proteins in the initial seed protein set. In Table 1, we show GO categories that are significantly enriched or impoverished in the sub-network. The 17 GO categories are filtered from 70 GO categories (data not shown) where available sub-network proteins has GO annotations. The filter criteria are 1) P-value over- or under- representation must be within 0.05 and 2) the total category count of GO in the whole network is greater than 10. In GO_TERM column, we listed three types of information: level 3 GO terms, GO term category type ('C' for cellular component, 'F' for molecular function, and 'P' for biological process; in parenthesis preceding the dash), and GO identifier (seven digit number following the dash in parenthesis). In the ENRICHMENT column, we listed two types of counts of proteins with GO annotation levels falling in the corresponding category: within core sub-network and whole network (in parenthesis). In the PVALUE column, we listed two numbers: p-value from significance test whether there is an over- or an under-representation (two numbers separated by a '/') of observed GO term category count in the sub-network. In the last CONCLUSION column, we used symbols to help us memorize test results: '++' to suggest significant over-representation when false discovery rate (FDR) controlled at 0.05, '--' to suggest significant under-representation when FDR controlled at 0.05, '+'

Table 1. Summarized Result for Observed Proteomics-level Changes in its Sub-network Context while Comparing Cisplatin-Resistant with Cisplatin-Sensitive Ovarian Cancer Cells. Only statistically significant changes are shown (see text for explanations of details and table formats). The changes include molecular function (F), biological processes (P), and cellular components (C).

		PVALULE	
GO_TERM	*ENRICHMENT*	*(OVER/UNDER)*	*CONCLUSION*
proton-transporting ATP synthase complex (C-0045259)	*8 (56)*	*0/1*	++
proton-transporting two-sector ATPase complex (C-0016469)	*4 (22)*	*.0001/1*	++
proteasome complex (sensu Eukaryota) (C-0000502)	*4 (66)*	*.0079/.9989*	+
organelle lumen (C-0043233)	*2 (13)*	*.0101/.9996*	+
myosin (C-0016459)	*2 (18)*	*.0191/.9988*	+
membrane (C-0016020)	*5 (1848)*	*1/0*	--
protein binding (F-0005515)	*31 (1412)*	*.0004/.9998*	++
drug binding (F-0008144)	*2 (13)*	*.0101/.9996*	+
isomerase activity (F-0016853)	*3 (41)*	*.0127/.9986*	+
transferase activity (F-0016740)	*8 (338)*	*.0493/.9802*	+
receptor binding (F-0005102)	*2 (944)*	*.9999/.0006*	-
metabolism (P-0008152)	*13 (556)*	*.0152/.9936*	+
regulation of viral life cycle (P-0050792)	*2 (20)*	*.0234/.9984*	+
cellular physiological process (P-0050875)	*64 (4496)*	*.0353/.9768*	+
detection of stimulus (P-0051606)	*2 (30)*	*.0496/.9947*	+
cell communication (P-0007154)	*15 (1881)*	*.9747/.0452*	-
organismal physiological process (P-0050874)	*5 (928)*	*.9892/.0289*	-

to suggest insignificant over-representation when FDR controlled at 0.05 but significant overrepresentation at native p-value=0.05, '-' to suggest insignificant over-representation when FDR controlled at 0.05 but significant overrepresentation at native p-value=0.05.

From the above table, we can obtain the following insights. First, there are abnormally high level of proton-transporting ATP synthase and ATPase production in the cell, suggesting unusually high oxidative energy production capability among cancerous function in cisplatin resistance cell lines over cisplatin sensitive cell lines. Second, although the protein interaction network is inherently enriched with proteins with "protein binding" capabilities (note 1412 proteins in the category from the whole network), the cisplatin-resistant cell line demonstrated unusually high level of protein-binding activities. This suggests that intracellular signaling cascades, not intercellular signaling (note the under-representation in "cell communication" category), is positively correlated with cisplatin-resistance. Third, the data suggest that the location of the biological activities of cisplatin resistant response take place in cytoplasm or nucleus, rather than on "membrane". This analysis gives essential clues to the overall picture of molecular signaling events for cisplatin resistant cell lines. We also obtained additional

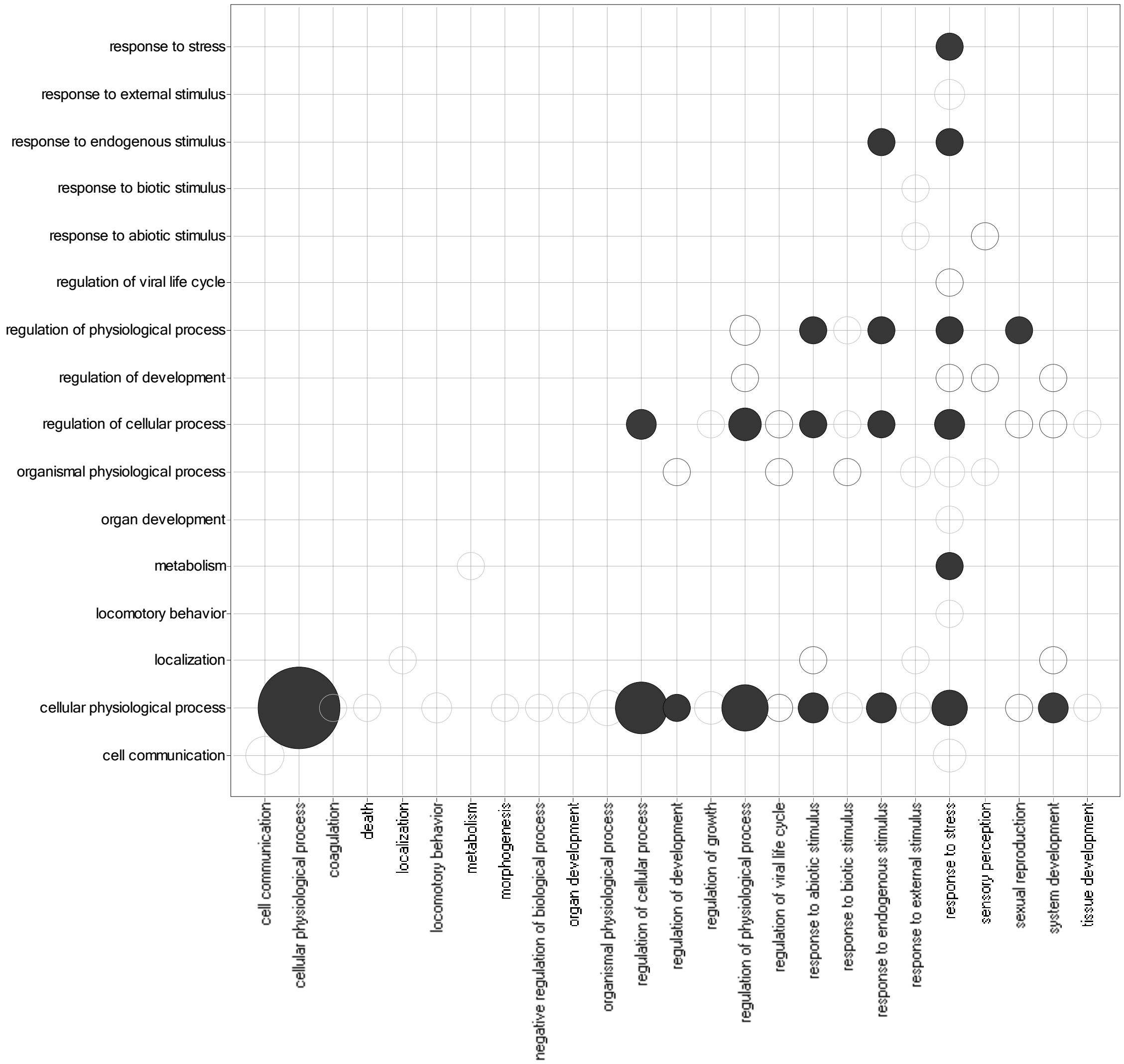

Figure 3. Cross-talk between related biological processes in Cisplatin-resistant Cell lines.

categorical enrichment data from different GO levels (not shown here due to space constraints).

3.3. Functional Category Cross-talks

We developed a 2-dimensional visualization matrix (extended from our technique described in [20]) to show significant cross-talks between GO categories in Fig. 3 (only biological processes at level=3 is shown due to space constraints). The size of node is inversely proportional to the p-value of interacting categories. The color legends are: red (dark) for interacting categories that are significant when FDR controlled at 0.05; and gray (light) colors for interacting categories that are not significant when FDR controlled at 0.05. The figure reveals additional interesting findings. First, cellular physiological processes are significantly activated in drug-resistant cell lines (the largest and reddest dot, at the bottom left corner). This could lead to further drill-down of protein interaction in the interacting category for biological validations (preliminary results; not shown). Second, these cellular physiological processes seem to be quite selective rather than comprehensive. For example, when looking at significant regulation of cellular response categories, significant cross-talk functional patterns strongly suggest the cellular and physiological responses arise from endogeneous, abiotic, and stress-related signals (internalized cisplatin causing DNA damage and inducing cell stress). Using a cross-talk matrix such as this, cancer biologists can quickly filter out other insignificant secondary responses (such as cell growth, cell development shown) to establish new prioritized hypothesis to test.

3.4. Visualization of the Activated Interaction Functional Sub-network

In Fig. 4, we show a visualization of activated biological process functional network, using a recently developed software tool "ProteoLens" [17]. ProteoLens is a biological network data mining and annotation

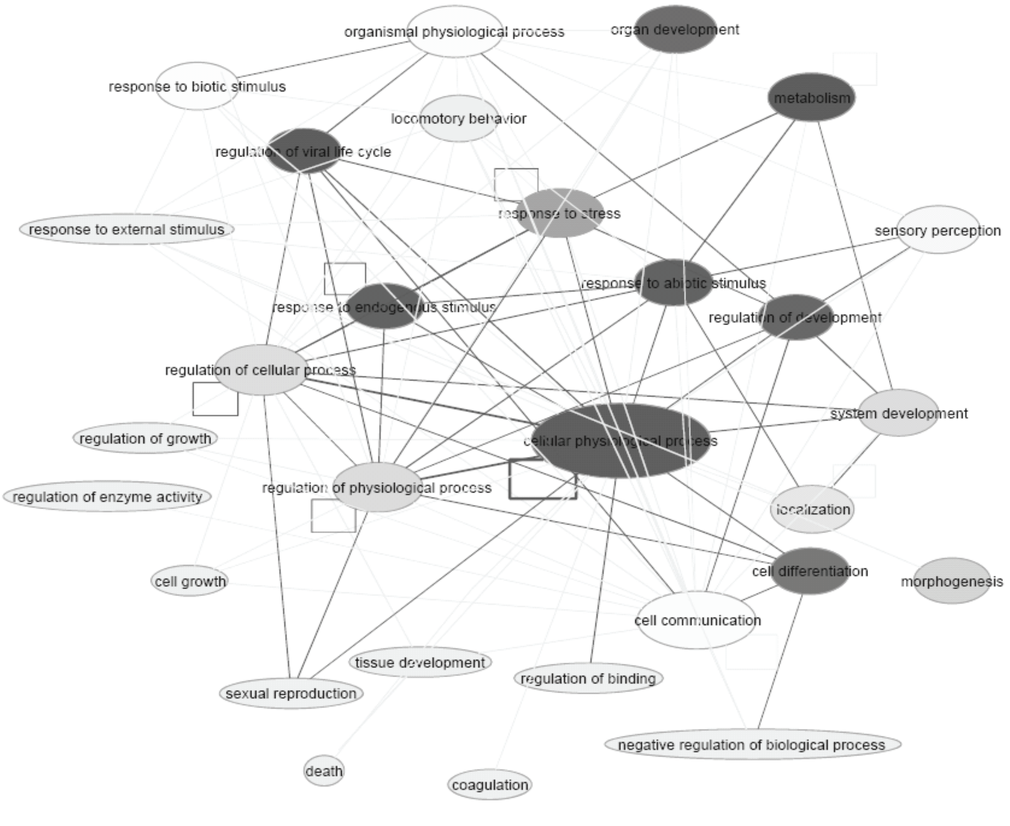
organismal physiological process
organ development
response to biotic stimulus
metabolism
locomotory behavior
regulation of viral life cycle
response to stress
response to external stimulus
sensory perception
response to abiotic stimulus
response to endogenous stimulus
regulation of development
regulation of cellular process
system development
regulation of growth
cellular physiological process
regulation of enzyme activity
regulation of physiological process
localization
cell differentiation
morphogenesis
cell growth
cell communication
tissue development
regulation of binding
sexual reproduction
negative regulation of biological process
death
coagulation

Figure 4. Overview of activated biological process functional network in cisplastin-resistant ovarian cancer cells.

platform, which supports standard GML files and relational data in the Oracle Database Management System (for additional details, visit http://www.proteolens.org/). In the figure, in contrast with regular protein interaction network, we encode nodes as significantly over-/under- represented protein functional categories, and edges as significantly interacting protein functional categories. Several additional information types are also represented. The p-values of interaction categories are inversely proportional to the thickness of edges, while the FDR=0.05 adjusted **interacting category significance** Boolean flags are indicated by line color: red (dark) for "significant" and blue (light) for "not significant". The original abundance (by count) of each functional category is encoded into node size. The p-values of activated **protein category significance** in the sub-network is encode as node color intensity, on a scale from light yellow (less significant) to dark red (more significant).

The resulting categorical network is both novel and informative. First, we confirm that cisplatin-resistant ovarian cancer cells demonstrated significant cellular changes in cell's overall physiological processes. These processes are first and foremost connected to cancer cell's native response to stimulus that is endogenous, abiotic, and stress-related as opposed to exogeneous, biotic, and related to tissue development. Second, while cell communication and cell growth are generally important to tumor mechanisms, we observe that molecular regulation of cell differentiation and development seems to be more relevant to the acquisition of drug resistance based on examination of this activated functional category interaction sub-network. Third, interestingly, we observe that the regulation of viral life cycle also plays very significant roles in the entire drug resistant process. This unknown observation may be further examined at protein levels to formulate hypothesis about acquired cisplatin resistance in ovarian cancer.

4. DISCUSSION

In this study, we showed that the key to interpreting Omics data is a systems biology approach, which is both "hypothesis-driven and data-driven, with the ultimate goal of integrating multi-dimensional biological signals at molecular signaling network levels. This is essential in unleashing the powerful potential for Proteomics techniques, which become a very powerful and efficient methodology in recent years for analysis of thousands of proteins on the basis of differences in their expression levels and post-translational modifications. In a systems biology approach, we integrate data from existing Omics databases and assemble different statistic as different visual cues in intuitive information visualization platforms. Our use of the functional 2-dimensional matrix and interaction category network is innovative. All these contributed to biological observations, which clearly demonstrate that cellular responses to genomic stress, in this case, DNA-damaging agent, are tightly associated with cisplatin resistance. Evidence for molecular regulation of cell differentiation and development also provide insight into the underlying mechanisms of cisplatin–resistance in ovarian cancer cells. Further examination of filtered subset of proteins in significantly detected functional categories and their cross-talks provide opportunities in modulating proteins involved in these biological processes/pathways to re-sensitize cisplatin resistant ovarian cancer cells.

We plan to conduct further studies to generate ranked list of proteins that significantly participated in the overall biological process of ovarian drug resistance as a group. We believe that the prioritized validation of these proteins in subsequent steps will further move us closer to finding molecular mechanism(s) causing ovarian cancer cells to become resistant to platinum-based chemotherapy. Meanwhile, we plan to collect Microarray data and conduct similar experiments to find out differences between the "Omics" results. We also plan to apply the systems biology approaches in different biological application domains.

Acknowledgments

This work was in part supported by a grant provided by Indiana University Purdue University Indianapolis to Dr. Jake Chen and by computer systems obtained by Indiana University through its relationship with Sun Microsystems Inc. as a Sun Center of Excellence. We thank Stephanie Burks for maintaining the high-end Sun servers and Oracle 10g servers. We also thank Jason Sisk and Kimberly Melluck from Indiana University School of Informatics for maintaining windows and Linux client computers, on which this study was partially conducted.

References

1. Yamamoto, K., Okamoto, A., Isonishi, S., Ochiai, K. and Ohtake, Y. (2001) Heat shock protein 27 was up-regulated in cisplatin resistant human ovarian tumor cell line and associated with the cisplatin resistance. *Cancer Lett,* **168,** 173-81.
2. Zamble, D.B. and Lippard, S.J. (1995) Cisplatin and DNA repair in cancer chemotherapy. *Trends Biochem Sci,* **20,** 435-9.
3. Auersperg, N., Edelson, M.I., Mok, S.C., Johnson, S.W. and Hamilton, T.C. (1998) The biology of ovarian cancer. *Semin Oncol,* **25,** 281-304.
4. Johnson, S.W., Laub, P.B., Beesley, J.S., Ozols, R.F. and Hamilton, T.C. (1997) Increased platinum-DNA damage tolerance is associated with cisplatin resistance and cross-resistance to various chemotherapeutic agents in unrelated human ovarian cancer cell lines. *Cancer Res,* **57,** 850-6.
5. Sakamoto, M., Kondo, A., Kawasaki, K., Goto, T., Sakamoto, H., Miyake, K., Koyamatsu, Y., Akiya, T., Iwabuchi, H., Muroya, T. *et al.* (2001) Analysis of gene expression profiles associated with cisplatin resistance in human ovarian cancer cell lines and tissues using cDNA microarray. *Hum Cell,* **14,** 305-15.
6. Chen, G., Gharib, T.G., Huang, C.C., Taylor, J.M., Misek, D.E., Kardia, S.L., Giordano, T.J., Iannettoni, M.D., Orringer, M.B., Hanash, S.M. *et al.* (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics,* **1,** 304-13.
7. Bhardwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics,* **21,** 2730-8.
8. Kitano, H. (2002) Systems biology: a brief overview. *Science,* **295,** 1662-4.
9. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* **25,** 25-9.
10. Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics,* **21,** 2076-82.
11. Pinto, F.R., Cowart, L.A., Hannun, Y.A., Rohrer, B. and Almeida, J.S. (2005) Local correlation of expression profiles with gene annotations--proof of concept for a general conciliatory method. *Bioinformatics,* **21,** 1037-45.
12. Higgs, R.E., Knierman, M.D., Gelfanova, V., Butler, J.P. and Hale, J.E. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res,* **4,** 1442-50.
13. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* **19,** 185-93.
14. Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics,* **4,** 1985-8.
15. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res,* **32,** D115-9.
16. Chen, J.Y., Shen, C. and Sivachenko, A. (2006) Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data. *Pacific Symposium on Biocomputing '06.* Maui, HI, Vol. 11, pp. 367-378.
17. Sivachenko, A., Chen, J. and Martin, C. (2005) ProteoLens: A Visual Data Mining Platform for Exploring Biological Networks (submitted). *Bioinformatics*.
18. Chen, J.Y., Pinkerton, S.L., Shen, C. and Wang, M. (2006) An Integrated Computational Proteomics Method to Extract Protein Targets for Fanconi Anemia Studies. *21st Annual ACM Symposium on Applied Computing.* Dijon, France, Vol. 1, pp. 173-179.
19. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.,* **29,** 1165–1188.
20. Chen, J.Y., Sivachenko, A.Y., Bell, R., Kurschner, C., Ota, I. and Sahasrabudhe, S. (2003) Initial Large-scale Exploration of Protein-protein Interactions in Human Brain. *IEEE Computer Society Bioinformatics 2003.* IEEE Computer Society Press, Stanford, California, pp. 229-234.