MINING MOLECULAR CONTEXTS OF CANCER VIA IN-SILICO CONDITIONING

Seungchan Kim[*] and Ina Sen

School of Computing and Informatics, Arizona State University,
Tempe, Arizona 85281,USA
[*]*Email: dolchan@asu.edu, ina.sen@asu.edu*

Micheal Bittner

Translational Genomics Research Institute,
Pheonix, Arizona 85281, USA
Email: mbittner@tgen.org

Cell maintains its specific status by tightly regulating a set of genes through various regulatory mechanisms. If there are aberrations that force cell to adjust its regulatory machinery away from the normal state to reliably provide proliferative signals and abrogate normal safeguards, it must achieve a new regulatory state different from the normal. Due to this tightly coordinated regulation, the expression of genes should show consistent patterns within a cellular context, for example, a subtype of tumor, but the behavior of those genes outside the context would rather become less consistent. Based on this hypothesis, we propose a method to identify genes whose expression pattern is significantly more consistent within a specific biological context, and also provide an algorithm to identify novel cellular contexts. The method was applied to previously published data sets to find possible novel biological contexts in conjunction with available clinical or drug sensitivity data. The software is currently written in Java and is available upon request from the corresponding author[*].

1. INTRODUCTION

The cellular system is very complex, arising from the interaction of many cellular components and processes. In order to maintain a specific state, the cell needs to tightly regulate various components using a host of regulatory mechanisms. A series of disruptions to the regulatory mechanisms, erodes the normal controls over proliferation, and produces a variety of other regulatory variations leading the cell to assume a significantly different state than its prior normal state, such as cancers.[14,15] To transition from normal to abnormal, (e.g. healthy to tumor), the functioning of the regulatory mechanism of the cellular system must be altered in significant ways. Such a change would result in an alteration of the way in which the cellular system interprets and acts upon certain kinds of input, in other words, a change of *cellular context*. While governing regulatory mechanism of normal context is disturbed in cancer, the persistent growth of the cancer implies that these cells retain a complex, reliable regulatory system capable of maintaining the enormous order required for the cell to live. The tumor's new behaviors now require a regulatory mechanism, possibly a different one from the regulatory mechanism that maintained the normal cell from which the tumor originated.

While many association-based approaches[4,12] have proven useful, one must look among all of the associated genes and attempt to group them on the basis of prior knowledge about the activities of the individual genes to identify particular processes. As the tool tries to look for more specific relationships among genes, it can find smaller groups of interacting genes, defined by the kinds of behaviors that arise from the way in which transcriptional regulation operates, improving the likelihood that such sets do represent interpretable hypothesis. More intriguingly, when the contextual information is unknown a priori, which is not unusual, capturing this implicit situational information, i.e. *cellular context*, based on observational evidence, and identifying genes with behavior specific to the context is a critical step toward the understanding of interactions among the participants and the discovery of its regulatory mechanism.

Recently, Segal et al developed the algorithms employing similar concepts[32] and applied those to a

[*] Corresponding author.

Saccharomyces cerevisiae expression data set to identify regulatory modules and their condition-specific regulators from gene expression data.[31] They also applied the method to perform an integrated analysis of 1,975 published microarrays spanning 22 tumor types to develop cancer module maps.[30] This method starts from initial partitions generated from clustering and utilizes prior biological knowledge such as Gene Ontology[3], KEGG (Kyoto Encyclopedia of Genes and Genomes)[20] and Gene MicroArray Pathway Profiler, if available, in that study. Our method does not explicitly depend on such knowledge but solely depends on data.

Biclustering[8] and Signature Algorithm[16,17] are two other methods comparable to the proposed method, which try to identify subsets of genes and samples. However, our method is inspired by the biologically interpretable master-slave model and has an inherent directionality in place, i.e. influence of master over the slave genes. Biclustering considers coherent gene-sample patterns but struggles with evaluating the separation between the identified biclusters, making its output not as easily interpretable. The signature algorithm on the other hand requires an initial seed gene list and builds the consistent condition list and gene list iteratively to identify transcription modules. The necessity for initial gene list limits the exploratory power of this algorithm. As the algorithm proceeds, dependant upon the genes/conditions included in progressive iterations, it may allow convergence to a separate module altogether, thereby losing the signal present in the initial list. Our method, context miner, identifies each context with a corresponding master gene and set of samples thereby ensuring the identification of a unique context and evaluates its statistical significance.

In the following sections, we first describe the algorithm to identify a set of genes that appear tightly regulated within known cellular contexts. We then describe a method to explore molecular and clinical patterns to identify all cellular contexts with consistent patterns that are statistically, significantly different from the rest of the data set. Lastly, we present the analyses of previously published data set; melanoma with gene expression profiles and gene expression along with drug activity data of NCI 60 cell lines, and conclude with some discussions.

2. METHODS

2.1. Identification of cellular contexts

It is assumed that when a cell maintains a specific cellular context, for example, a phenotype, it tightly regulates a battery of genes, which would show rather deterministic transcriptional activities. When the cell moves away from this cellular context or changes to a different cellular state, the behavior of the same set of genes will not appear as deterministic since they now behave without control signals (intrinsic stochastic behavior) or each gene comes under the control of various other external controls.

In this section, we first describe novel statistics to identify a set of genes with more deterministic transcriptional behavior within a given cellular context than outside the context. Once a set of genes is identified, we evaluate the statistical significance of such a finding. While the algorithms are described and applied in the context of transcriptional activities, we later explain how to use the proposed method for gene expression data integrated with other types of data such as array-based comparative genomic hybridization (aCGH) data and other clinical parameters such as drug sensitivity.

2.2. Consistency statistics: interference and crosstalk

To identify a set of genes with consistent transcriptional behavior within a specific cellular context, we need statistics to evaluate consistency and/or inconsistency within and outside specified context. Here, we consider a context c to be given by specifying a subset of samples, S, assumed to share certain phenotypes resulting from being governed by common biological processes or regulatory mechanisms. We formulate the hypothesis as follows. Let us assume a cell can be in any of the different cellular statuses, $C \in (c_1, c_2, \ldots c_k)$. In other words, specifying context c_j will partition samples into two groups, one that would reflect the cellular status defining the context and the other that does not. For example, a clinical parameter such as drug responsiveness can be considered a conditioning factor, partitioning patients into two groups; one being responsive and the other not. The two statistical parameters, *interference* and *crosstalk*, can be used to determine whether a gene is being regulated within a given cellular context.[21] The interference[b], $\delta_k^{(j)}$, for a gene g_k given a cellular context c_j, is the extent to which latent variables (external controls sensitive but not specific to context) interfere with the regulatory signal from a master gene, $G^{(j)}$:

$$\delta_k^{(j)} = 1 - \Pr\left(g_k = ON \middle| C = c_j\right). \tag{1}$$

[b] 1 - $\delta_k^{(j)}$, has the same form of equation as the *precision*. However, the interference was motivated by biological insight about gene regulation and we will keep the term as is.

and the crosstalk $\eta_k^{(j)}$ is defined as the probability that the gene, g_k is being regulated (by external control), when the cellular context is not c_j:

$$\eta_k^{(j)} = \Pr\left(g_k = ON \middle| C \neq c_j\right) \qquad (2)$$

The equations above can be modified to consider g_k = *OFF* as well.

A gene is determined to be specific to the given context if both interference and crosstalk are significantly low. Given a subtype of tumor, for example, we identify genes with significantly low interference and crosstalk as being tightly regulated within the tumor (See Fig. 1). For example, we want to find a set of genes that are consistently up-regulated (ON) only within a group of patients who respond to a therapy but not outside. The interference and the crosstalk can be used to find such genes. This approach is different from typical t-test where a gene needs to be differentially expressed (ON in one group and OFF in the other group). The interference and crosstalk allow certain level of up-regulation outside the context as long as it is not as consistent as within the context.

Since both the interference and crosstalk parameters are estimated from the observations, the statistical significance of the estimated values should be considered in order to avoid highly possible false positives or false discoveries. Let N be the number of observations made and assume that there are two different classes (different subtypes of diseases or prognosis). For a gene, assume there are n_0 OFF status and n_1 (= N - n_0) ON status overall in the observations. When we partition the N samples into two groups based on their class labels, n in the first class and $N - n$ in the other, let $l \leq n_1$ ON samples get assigned to the first group. Let us denote the subset of samples associated with the first group by $S^{(+)}$ and the other by $S^{(-)}$. Using the Eqs. (1-2), we estimate both interference and crosstalk, but we would like to know the probability of obtaining those values by chance given the observations. Since we partition the samples to acquire those numbers, this probability is same as the probability that we partition the samples to have exactly the same configuration given the parameters above, (N, n_1, n, l), and this can be computed via hyper-geometric probability as follows in Eq. 3:

$$\Pr\left(L = l; N, n_1, n\right) = \binom{n_1}{l}\binom{N-n_1}{n-l} \Big/ \binom{N}{n} \qquad (3)$$

We then define the probability, given (N, n_1, n), that the gene is consistently expressed (ON) l times or more as:

$$\Pr\left(L \geq l; N, n_1, n\right) = \sum_{i=l}^{\min(n,n_1)} \Pr\left(L = i; N, n_1, n\right) \qquad (4)$$

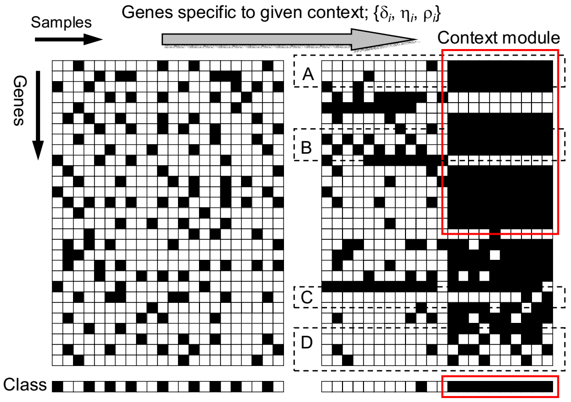

Figure 1. Context module: Group **A** indicates genes with both low cross-talk and low interference with statistical significance. **B** presents genes with low interference but high cross-talk, **C** with no statistical significance, and **D** with high interference or cross-talk. A set of genes identified in A and B is called context module.

As more ON's are assigned to the class of our interest, both the interference and the crosstalk parameters decrease. Therefore, Pr($L \geq l$; N, n_1, n) is the probability that those parameters are estimated at the same values or higher. If this probability is very low, such as less than 0.05, it is rare to find those estimated values by chance, *i.e.*, it is significantly different from what can be found by chance. For a given context, c_j and corresponding subsets $S_j^{(+)}$ and $S_j^{(-)}$, and a gene, g_k, with the parameter, (N, $n^{(k)}_1$, $n^{(k)}$, $l^{(k)}$), we denote the probability Pr($L \geq l^{(k)}$; N, $n^{(k)}_1$, $n^{(k)}$) by $\rho_k^{(j)}$.

The set of genes with low interference and crosstalk and yet statistically significant are identified to be specifically highly correlated within a given cellular context. The set of genes of interest to biologists focusing on subtypes of cancer are the ones with low interference (Eq. 1) within the subset of tissues from the corresponding subtype and low crosstalk (Eq. 2) outside the subset with low probability of finding such gene by chance (Eq. 4).

2.3. Interrogating contexts via in-silico conditioning

In practice, such explicit knowledge about contexts as clinical parameters is often not known a priori. In this section, we describe a method to systematically identify possible molecular contexts. The premise is that each context is conditioned by a gene, i.e. master, or other external, cellular conditioning parameters such as clinical parameters. The method interrogated each gene or clinical parameter if one of its states, for example, ON or OFF, could be interpreted as a conditioning factor. This was done by grouping samples into two: first with the samples where the state of conditioning parameter is set to a specific state, and the other with the rest of the samples. Then, by applying the method described in the previous section, we identified the genes seemingly tightly regulated in such conditions.

This is similar to biologists' manipulating the status of a gene or conditioning cells to investigate its down stream effect. Biologists often use ectopic expression[10,13,19] or gene silencing techniques such as RNA interference[5,11,23] to either increase or decrease the expression levels, respectively. In our case, it is done computationally after the data is collected. Thus, we call this *in-silico* conditioning.

Each conditioning yields a subset of samples, i.e. context, where a set of genes that appear tightly regulated within the context are obtained. Depending on the number of samples and the number of genes, the context might be statistically insufficiently distinctive; the pattern of the similar size of samples and genes might be found by chance. The next subsection addresses this issue.

2.4. Significance test for identified contexts

We assessed the probability of finding a context pattern where the same number of or more genes were tightly regulated across same number of samples by chance. Let (M, N) denote data size where M is the total number genes and N is the number of samples in data set. We also let m and n denote the number of co-regulated genes and the number of observations in an identified context, respectively. We estimated Pr($m' \geq m \mid n' = n$), the probability that a context regulates larger or equal number of genes than m, given the sub-sample size n. This probability was estimated via re-sampling method. More specifically, we randomly split given data set into two groups of which the one was of sample size n (context candidate) and the other of $N - n$. We then applied the same set of statistics (Eqs. 1-2) to identify the number of genes filtered by the same thresholds for interference (η), crosstalk (δ) and p-value (ρ). By repeating this procedure many times, we estimated Pr($m' \geq m \mid n' = n$). The accuracy of the estimation is based on the number of repetitions. In typical setting, no less than 1,000 repetitions were required to provide distribution with enough statistical power. Using this re-sampling-based approach, we could assess the statistical significance of identified contexts and consider only significant patterns for further analysis.

2.5. Data quantization

To use the proposed method, gene expression data needs to be quantized. If the data is pre-quantized by a sophisticated method such as described in Chen et al.[6,7] we use them as is. If not, there are several other methods available: fold-changes, heuristic-based,[9,33] and model-based approaches.[35,37] While relevant, the discussion of the quantization issue is beyond the scope of this study, we therefore leave the further discussion to those studies.

2.6. Data quantization

To use the proposed method, gene expression data needs to be quantized. If the data is pre-quantized by a sophisticated method such as described in Chen et al.[6,7] we use them as is. If not, there are several other methods available: fold-changes, heuristic-based,[9,33] and model-based approaches.[35,37] While relevant, the discussion of the quantization issue is beyond the scope of this study, we therefore leave the further discussion to those studies.

3. EVALUATION OF THE ALGORITHM

To evaluate its utility and the performance of the proposed algorithm, we used simulation-based experiments. To generate the synthetic data, we started with a set of master-slave relations, which consisted of a master, a set of slaves and a set of rules between the master and the slaves. We then added the conditioning and crosstalk parameters to specify the strength of regulation from the master to the slaves, which added randomness to the relations. This master-slave relation defined a cellular context. Since typical cells have many such cellular mechanisms actively operating simultaneously, we specified multiple cellular contexts and samples were drawn, as measurements were made, from the set of cellular contexts. The process is illustrated in Figure. 2.

In Figure 2, the third sample, row (red[c] (R), black (B), B, green (G), R, B, G, R, G, R, R, G, G, B, G, B) has been drawn from the same cellular context (first graph) as the first sample, row (R, B, B, G, R, B, B, R, G, R, R, G, G, R, G, B), but because of the randomness introduced by the conditioning and crosstalk parameters, on being sampled they do not show the identical gene expression profile, which is also typical in real biological observations.

In the simulation done, we used four different cellular contexts, one master gene in each context, and

[c] "red" appears dark gray, and "green" appears light gray, due to the conversion to grayscale to comply with the conference guideline.

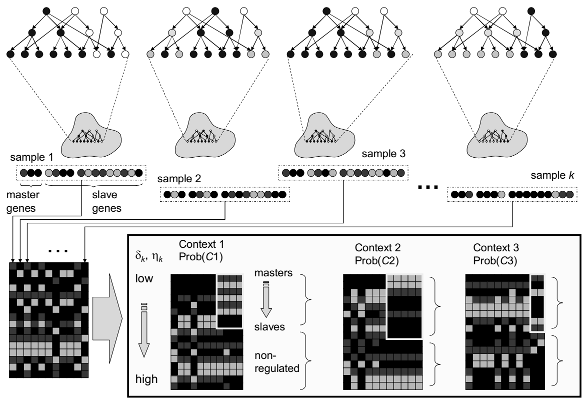

Figure 2. Master-Slave regulatory model to simulate gene expression data: different numbers of samples are drawn from each context (three contexts).

105 genes in a simulated data. The main focus of the simulation was to find the effects of number of samples drawn for each context and the number of regulated genes in each context, to the performance of the algorithm. For each data, four different contexts were sampled with different sampling rates of 5, 10, 15, and 25 observations. Also different numbers of slaves were tested for each context: 10, 15, 25, and 40 out of 105 total genes. Once the data was generated, for each cellular context, we tested how accurately the algorithm identified master and slaves corresponding to the cellular context. For comparison, other statistics-based method (correlation) and information-based method (mutual information) were also used to identify such sets. These methods have been popularly used in microarray analyses such as clustering to identify co-regulated genes within same cellular context. For the measures of performance, we used false positive (FP; the number of genes identified as being regulated that in fact are not), false negative (FN; the number of genes not identified as being regulated that in fact are), and total error (FP + FN). Each parameter combination was repeated 200 times to compute the average performance. The results are shown in Figure. 3 which compares the performance of correlation and mutual information with the context miner in terms of error.

As we can see, in all cases, the proposed algorithm (context) outperformed the other algorithms (mutual information and correlation). There was not much difference between the other two algorithms in terms of performance. Fig. 3 (a) shows the effect of the number of regulated genes in each context. It is shown that significant portion of total error comes from false negative (FN). It also shows that FN increases as the number of slaves increases while false positive (FP) remains unchanged, which is somewhat expected. Fig. 3 (b) shows the impact of the size of context (sample size). While FN remains relatively unchanged, FP decreases significantly as the sample size increases. Overall error for the proposed method also remains relatively low at 4 to 8% for different number of slaves in each context and 5 to 9% for different sample sizes. Therefore, the simulation study proves that the proposed method can be effectively used to identify a set of genes that are specifically regulated within a specific cellular context.

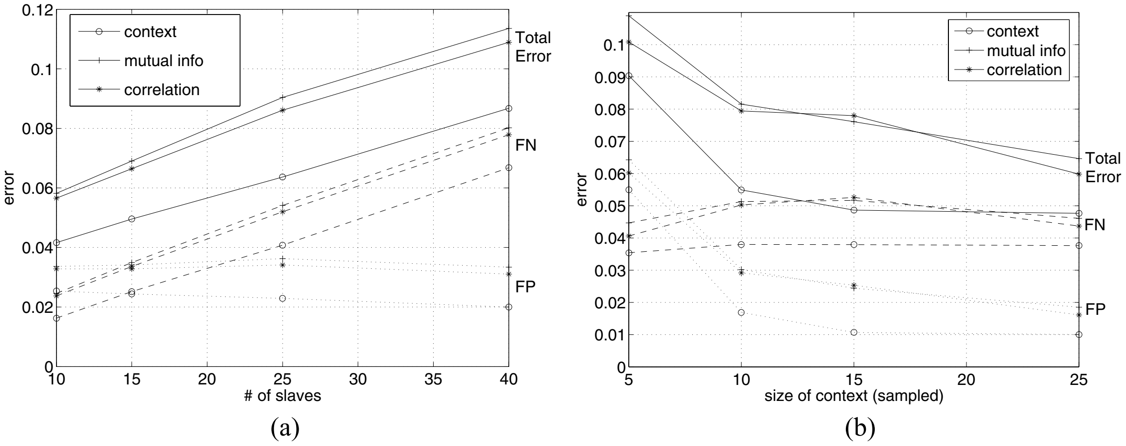

Figure 3. The comparisons of the proposed method against other association-based methods shows the proposed methods outperform the others for both increasing number of regulated genes (slaves) (a) and the size of contexts (samples) (b).

4. RESULTS

To show a proof of principle, we applied the method to a previously published melanoma data set[4] and gene expression data with drug activity data for the NCI 60 cell lines.[29] The latter will illustrate how multiple types of data (gene expression and drug activity data) can be combined in the analysis to identify interesting patterns of interactions not only among genes but also between genes and drugs.

4.1. Analysis of melanoma gene expression profile

Melanoma data set consists of 8,607 genes and 31 samples. After preliminary analysis and filtering according the method described in the original paper, we extracted 587 genes which were then used for this study. In the original study of melanoma, there existed very tightly clustered samples (major cluster) with less motility and invasiveness. We, therefore, first identified a set of genes that displayed consistent expression in the major cluster using our method. Top two genes identified in the original paper, WNT5A and MLANA, as well as SNCA and EDNRB were found at the top of our list (data not shown). Also, there were some new genes identified with high consistency, which are interesting candidates for further study.

Another finding in the original paper was the regulatory control of Wnt5a when it is *highly expressed*. Thus, the samples conditioned by a high expression of WNT5A were found to be strongly associated with the high expressions of MLANA, DKK3, SERPINE1, MT1X, KAI1, BRD2, and TRAM1. The involvement of these genes with melanoma development is unknown but the consistency of their transcriptional activities warrants further investigation. However, the regulation of MLANA by WNT5A has been recently reported.[34]

To unravel novel molecular contexts related to melanoma, we applied the algorithm as described in section 2. Using more than 10,000 re-sampling, we identified more than 100 contexts with $p < 0.005$. Table 1 lists the contexts with $p < 1e\text{-}4$.

In Table 1, note that two distinct states of MLANA make up two different contexts; the first one is when it is normal and has 17 genes (excluding itself) being regulated, and the other is when it is up-regulated and has 14 genes (excluding itself) as regulated. When compared, two contexts share only one gene (MYLK) in common as regulated, but in distinct states. This confirms our assumption that a gene can be regulated by different regulators (masters) when cellular context changes. Further investigation revealed that more than 100 contexts identified with $p < 0.005$ can be grouped into less than 30 larger contexts with unique hierarchical structures (data not shown), upregulation of WNT5A and upregulation of JUN being among them. Upregulation of WNT5A is interesting because of its regulation of MLANA, as reported in Weeraratna et al.[34] Upregulation of JUN became also interesting because of our biologist's other supporting biological evidences. These two contexts are exclusive, implying two distinctive molecular contexts relevant to melanoma.

Table 1. Cellular contexts (denoted by masters) identified with the statistical significance, p < 1e-4.

Conditioners	State	m	n	Pr(m′>m|n′=n)
MLANA		10	18	0.0000992
PLP1	+	23	7	0.0000992
MLANA	+	21	15	0.0000993
FBN2		13	12	0.0000994
MMP3	-	12	14	0.0000994
TCEB3	-	17	6	0.0000994
LOC646762	+	9	21	0.0000994
MYLK	+	20	16	0.0000995
DUSP1	+	7	37	0.0000995
MMP3		16	11	0.0000995
IFIT1		14	15	0.0000995
MBP		15	8	0.0000995
EDNRB		6	43	0.0000995
SNED1	-	4	68	0.0000996
WNT5A	-	24	4	0.0000997

Conditioners are the genes used to *in-silico* condition context. State indicates the expression states of corresponding conditioner. *m* and *n* denote the number of samples where the conditioner is kept at the state and the number of regulated genes (including the conditioner itself) within such context. The last column shows the probability that a context with equal or larger size can be identified by chance using re-sampling method.

Ongoing work to elucidate the effects of WNT5A in melanoma has revealed that high levels of WNT5A in melanocytes are associated with higher production of the cytokine IL6. Work from a number of laboratories shows that the transcription factor Mitf, which drives Melan-A transcription is itself regulated by Pax3 and Sox10,[25] and that this regulatory pathway can be inhibited in melanoma by IL6 stimulation, which affects Pax3.[19]

In the other context considered, the stimulation of transcription of DUSP1 by JUN is seen. DUSP1 expression is known to be upregulated by the onset of chronic proliferation[24], and is known to inactivate

MAPK through dephosphorylation.[1] DUSP1 transcription is activated in a wide variety of stresses, and the exact transcription factors involved are not well worked out.[22] It is likely that the high predictivity of DUSP1 transcription by increased JUN transcription arises from the simultaneous stimulation of both JUN and DUSP1 that occurs when cells go into chronic proliferative states. Interestingly, recent study found the possibility of potential diagnostic relevance of DUSP expression in tumors.[2]

4.2. Analysis of NCI 60 cell line gene expression and drug sensitivity data

We extended the concept of finding conditioning factors from only genes, to elements which influence, regulate or act specific to the existing cellular state. Any such factor would also be bound by the constraints in place due to cellular state. Applying our method to such disparate datasets such as aCGH, gene expression and/or drug activity data, would allow us to witness the possible underlying patterns of the inter- and intra-relationships in them. Using the NCI60 data set we show that the contexts identified can help guide further studies of drug effectiveness and mechanism of action.

4.2.1. *Data preparation*

To provide an example of exploratory functions possible by our method, we applied it to the NCI 60 drug data. The dataset consisted of the drug activity data of 118 drugs and the gene expression data of 1375 genes across the NCI-60 cell lines.[29] The original paper related this data to sensitivity to therapy rather than to molecular consequences of the therapy, as the gene expression patterns were determined in untreated cells.

The drug activity was represented in a matrix with –log GI_{50} values, where GI_{50} is an indicator of the growth inhibition by the compound on the cell line.The application of our method can be summarized into three steps - scaling of data to comparable form (normalization), combining these forms, and applying our method to obtain contexts corresponding to the different conditioning factors. In order to scale the different data sources, the drug matrix was normalized by subtracting its row-wise mean and dividing by its row-wise standard deviation. For the gene expression matrix, no transformation was applied, the matrix being already normalized. Next, matrix entries were quantized on the basis of two-fold changes, for statistical significance. Then both quantized matrices were combined into one and used as the input data for the context analysis. The generalized process is captured in Figure 4.

4.2.2. *Patterns of drug-gene interactions*

The context analysis on the NCI 60 drug activity and gene expression data resulted in 4153 contexts. Among those, we focused on the contexts where at least one drug and a gene were included, which resulted in 243 contexts. At last, only 27 contexts were found to be statistically significant with p-value less than 0.01,

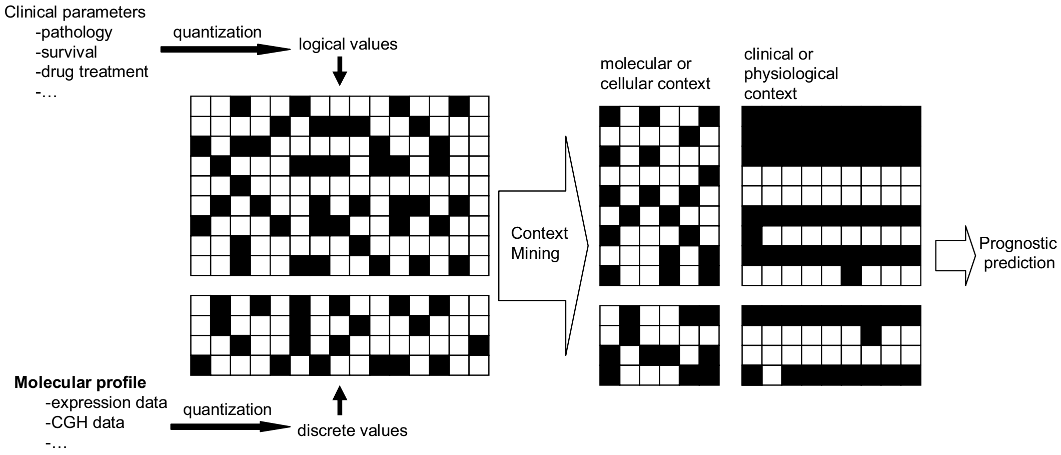

Figure 4. Combining genomic data and clinical parameters to identify cellular contexts - Drug activity data and gene expression data from NCI gene expression database NCI 60 cell lines was discretized independently into ternary values and then combined into a single data set. Using each drug or gene as the conditioning factor, context analysis was carried out to obtain contexts focusing on drug-gene combinations.

Table 2. Top 27 contexts identified from combined drug data and gene expression data, with statistical significance p<0.01. The first column represents the conditioning factors (gene/drug/disease) of the context. Genes are represented by gene symbols. In case of drugs, the drug name is shown with the mechanism of action e.g. [TU]. The second, third and fourth columns lists the number of conditioning factors, number of cell lines and total number of drug/gene elements respectively identified for the context. The fifth column reports the p-value of finding such a context. The final column reports the number of drugs that were found to be highly active in that context.

	SW	S	G	Pr(G+|S)	Drugs
Gene					
PTK2	2	4	184	0.00163	33
RAB7	1	5	132	0.00163	0
GJA4	2	3	241	0.00169	39
HEXB	1	2	200	0.00172	37
MMP14	1	3	173	0.00253	5
TWF1	1	5	102	0.00327	3
CORO1A	1	5	102	0.00327	5
TDG	1	3	164	0.00337	16
GLUL	1	3	145	0.00422	3
ISGF3G	4	4	159	0.00489	3
KCNQ4	2	5	93	0.00490	4
-	1	3	139	0.00506	9
-	1	2	174	0.00517	44
MYL3	2	4	145	0.00653	9
RP6-213H19.1	1	4	120	0.00734	3
MAPRE2	2	3	118	0.00759	3
KLF6	1	3	118	0.00759	12
-	1	4	117	0.00816	5
-	1	3	114	0.00843	5
IRX3	1	4	107	0.00897	2
REEP5	3	2	144	0.00948	6
-	1	4	100	0.00979	0
Drugs					
7-Epi-10-deacetylbaccatin III [TU]	1	2	190	0.00259	7
Camptothecin,20-ester (S) [T1]	1	42	7	0.00382	0
Camptothecin,11-HOMe (RS) [T1]	1	2	160	0.00603	42
Taxol analog [TU]	1	54	4	0.00771	0
Disease					
Leukemia	2	6	78	0.00485	2

[TU] is Tubulin–active antimitotic agents and [T1] – Topoisomerase I inhibitor.

which are displayed in Table 2. Among these identified contexts, we proposed how these elements act with each other, based on the domain knowledge, annotations or functionality.

We observed that the majority of the contexts reflected patterns found in the original paper.[29] For example, the two breast cancer cell lines positive for oestrogen receptor, T-47D and MCF7, clustered together in the original paper, were also found to be grouped together in our analysis. The context identified showed higher activity of drug 11-formyl Camptothecin (RS) than its counterpart Camptothecin, 11-HOMe(RS).

For the two cell lines (MDA-MB-435 and MDA-N), there were two filtered contexts of interest In the first context with only these two cell lines grouped, drug 7-Epi-10-deacetylbaccatin III (Taxol Analog NSC No.

656178), Paclitaxel and other Taxol analog drugs with the mechanism of action as Tubulin-active antimitotic agents (TU) displayed highly active status. In the second context, conditioned by gene RAB7, these two cell lines were grouped together with Melanoma cell lines (MALME-3M, SK-MEL-5 and UACC-62).

Interestingly, the drugs identified in this context as being consistent were Cyclocytidine and Cyctarabine(araC), belonging to DNA synthesis inhibitor mechanism (Ds). However, they did not display high activity across all these samples, while Taxol analog drugs were highly active in these two breast cancer cell lines. In the original paper, MDA-MB-435 and MDA-N cell lines clustered closely with Melanoma cell lines.[29] The authors had discussed that the MDA-MB-435 and its Erb/B2 transfectant MDA-N expressed large number of genes characteristic of melanoma, and recent findings now group these two as a subtype of Melanoma itself.[26-28] However, the finding in our study may indicate that they still do not use the same mechanisms in drug responses.

In Table 2, many of the contexts include drugs that have different mechanism of action. Every context depicts the common transcriptional activities of given cell lines, for example, subtypes of cancers with shared transcriptional behavior. It is possible that in order to stop proliferation of the cell, different points of the regulatory mechanisms present in cancer cells are targeted. Thus depending upon drug target point, varying degree of potency of drug would be established, effective in arresting the cancer development.

Our initial purpose of being able to attribute the drug to a particular mechanism seemed thwarted by the inclusion of drug in multiple contexts, showing more than one type of mechanisms active in each context. Considering the previous argument, we tried to improve the prediction of mechanism of action of drug by finding maximum overlap between biological processes (GO terms) of the genes targeted by drug with unknown action and those of drugs with known action. Greater overlap would imply similarity in mechanism of action.

We tried assigning the mechanism of action of drug Inosine-glycodialdehyde (Inox) by studying other drugs in all contexts which include Inox. In the context conditioned by IRX3, Inox showed similar activity to 11-Formyl-20(RS)-Camptothecin, of mechanism T1, topoisomerase 1 inhibitor. In the context conditioned by gene TWF1, it showed high activity along with drugs Dichloroallyl-lawsone and Pyrazofurin of mechanism Rs, RNA synthesis inhibitor. This context consisted of Leukemia cell lines CCRF-CEM, K-562, MOLT-4, HL-60 and RPMI-8226.

We extracted for each drug the corresponding target genes from PubGene[18] and ran the obtained lists through GoMiner.[36] On matching the significant GO terms (with p-value<0.05) we found that although there were less than 10 exact matches but the terms displayed more coherency in terms of function to Rs mechanism derived GO terms. For Inosine-glycodialdehyde, we found GO:0000122, GO:0045892 which relate to negative regulation of transcription (from RNA polymerase II promoter and DNA-dependant). GO terms matching those from Pyrazofurin and dichloroallyl-lawsone (Rs mechanism) included GO:0006220, GO:0009058, GO:0009165 and GO:0044249, related to nucleotide metabolism and biosysnthesis. There was no significant GO term match between those derived from Inox and those from Camptothecin.

Some contexts group different cell lines possibly implying an underlying similarity in the regulatory mechanism in place, irrespective of the tissue of origin. This allows identification of drugs which could be potent in these particular cancer subtypes, allowing us to span and target a greater range of cancer types using the same drug. By finding targeted mechanisms by concentrating on annotations such as GO terms would allow greater power in our ability to prescribe effective drugs.

5. CONCLUSION

We propose a method to identify putative cellular contexts via in-silico conditioning, which, if applied to the study of cancer, could lead to the discovery of subtypes of the disease not obvious at the histological level but possibly explained at molecular levels and carry prognostic relevance. The method can be applied to the experimental data with disparate data sources to improve understanding of the multilayer interactivity of biological components and help direct further studies. We used this method on public datasets of melanoma and gene expression data with drug activity data of NCI 60 cell lines. In melanoma study, we identified distinctive transcriptional patterns, one of which can be of clinical importance. The contexts analyzed imply some concerted pattern amongst the different components, and may be necessary to allow integration of biological data using prior knowledge to guide the combination and comprehension of the data.

The current method is limited to only one conditioning parameter due to its exhaustive search which grows exponentially with the number of conditioning factors. Biological systems often require multivariate conditioning, and we are currently exploring extension of the algorithm to address it.

Acknowledgments

We thank Dr. Haiyong Han (Translational Genomics Research Institute) for providing help interpreting the results of NCI 60 cell line data, Mr. Verdicchio and Ms. Ramesh for helping us with data analysis. We also thank our reviewers for giving us invaluable feedback. This study is partially supported by ASU/Mayo Collaboration Initiative Grant (#016851-001), NIH PO1-CA27502-23, P01 CA109552-01A1 and U19 AI067773.

References

1. Alessi, D. R., C. Smythe, et al. The human CL100 gene encodes a Tyr/Thr-protein phosphatase which potently and specifically inactivates MAP kinase and suppresses its activation by oncogenic ras in Xenopus oocyte extracts. *Oncogene* 1993; **8**(7): 2015-20.
2. Amit, I., A. Citri, et al. A module of negative feedback regulators defines growth factor signaling. *Nat Genet* 2007 **accepted**.
3. Ashburner, M., C. A. Ball, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**(1): 25-9.
4. Bittner, M., P. Meltzer, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; **406**(6795): 536-40.
5. Caplen, N. J., S. Parrish, et al. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc Natl Acad Sci U S A* 2001; **98**(17): 9742-7.
6. Chen, Y., E. R. Dougherty, et al. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images." *Journal of Biomedical Optics* 1997; **2**: 364-74.
7. Chen, Y., V. Kamat, et al. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 2002; **18**(9): 1207-15.
8. Cheng, Y. and G. M. Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 2000; **8**: 93-103.
9. Chung, T.-H. and S. Kim. Quantization of Global Gene Expression Data. *Internationa Conference onl Machine Learning and Application (ICMLA)* 2006, FL.
10. Davidson, E. H., J. P. Rast, et al. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev Biol* 2002; **246**(1): 162-90.
11. Fire, A., S. Xu, et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998; **391**(6669): 806-11.
12. Golub, T. R., D. K. Slonim, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 1999; **286**: 531--537.
13. Hahn, W. C., C. M. Counter, et al. Creation of human tumour cells with defined genetic elements. *Nature* 1999; **400**(6743): 464-8.
14. Hahn, W. C. and R. A. Weinberg. Modelling the molecular circuitry of cancer. *Nat Rev Cancer* 2002; **2**(5): 331-41.
15. Hanahan, D. and R. A. Weinberg . The hallmarks of cancer. *Cell* 2000; **100**(1): 57-70.
16. Ihmels, J., S. Bergmann, et al. Defining transcription modules using large-scale gene expression data. *Bioinformatics* 2004; **20**(13): 1993-2003.
17. Ihmels, J., G. Friedlander, et al. Revealing modular organization in the yeast transcriptional network. *Nat Genet* 2002; **31**(4): 370-7.
18. Jenssen, T. K., A. Laegreid, et al. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001; **28**(1): 21-8.
19. Kamaraju, A. K., C. Bertolotto, et al. Pax3 down-regulation and shut-off of melanogenesis in melanoma B16/F10.9 by interleukin-6 receptor signaling. *J Biol Chem* 2002; **277**(17): 15132-41.
20. Kanehisa, M. and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**(1): 27-30.
21. Kim, S., E. R. Dougherty, et al. Cellular contexts from gene expression profile. *IEEE International Workshop on Genomic Signal Processing and Statistics* 2005, Newport, RI.
22. Li, M., J. Y. Zhou, et al. The phosphatase MKP1 is a transcriptional target of p53 involved in cell cycle regulation. *J Biol Chem* 2003; **278**(42): 41059-68.
23. Mousses, S., N. J. Caplen, et al. RNAi microarray analysis in cultured mammalian cells. *Genome Res* 2003; **13**(10): 2341-7.
24. Noguchi, T., R. Metz, et al. Structure, mapping, and expression of erp, a growth factor-inducible gene encoding a nontransmembrane protein tyrosine phosphatase, and effect of ERP on cell growth. *Mol Cell Biol* 1993; **13**(9): 5195-205.
25. Potterf, S. B., M. Furumura, et al. Transcription factor hierarchy in Waardenburg syndrome:

regulation of MITF expression by SOX10 and PAX3. *Hum Genet* 2000; **107**(1): 1-6.

26. Rae, J. M., C. J. Creighton, et al. MDA-MB-435 cells are derived from M14 Melanoma cells--a loss for breast cancer, but a boon for melanoma research. *Breast Cancer Res Treat* 2006.
27. Rae, J. M., S. J. Ramus, et al. Common origins of MDA-MB-435 cells from various sources with those shown to have melanoma properties. *Clin Exp Metastasis* 2004; **21**(6): 543-52.
28. Ross, D. T., U. Scherf, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000; **24**(3): 227-35.
29. Scherf, U., D. T. Ross, et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000; **24**(3): 236-44.
30. Segal, E., N. Friedman, et al. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 2004; **36**(10): 1090-8.
31. Segal, E., M. Shapira, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003; **34**(2): 166-76.
32. Segal, E., H. Wang, et al. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* 2003; **19 Suppl 1**: i264-71.
33. Shmulevich, I. and W. Zhang. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 2002; **18**(4): 555-65.
34. Weeraratna, A. T., Y. Jiang, et al. Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. *Cancer Cell* 2002; **1**(3): 279-88.
35. Yeung, K. Y., C. Fraley, et al. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001; **17**(10): 977-87.
36. Zeeberg, B. R., W. Feng, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003; **4**(4): R28.
37. Zhou, X., X. Wang, et al. Binarization of microarray data on the basis of a mixture model. *Mol Cancer Ther* 2003; **2**(7): 679-84.