



A novel method for generation of signature networks as biomarkers from complex high throughput data

Yuri Nikolsky, Sean Ekins*, Tatiana Nikolskaya, Andrej Bugrim

Computational Biology Genego Inc., 500 Renaissance Drive, Suite 106, St. Joseph, MI 49085, USA

Received 15 January 2005; received in revised form 13 February 2005; accepted 14 February 2005

Available online 24 May 2005

Abstract

Traditionally, gene signatures are statistically deduced from large gene expression and proteomics datasets and have been applied as an experimental molecular diagnostic technique that is sensitive to experimental design and statistical treatment. We have developed and applied the approach of “signature networks” which overcomes some of the drawbacks of clustering methods. We have demonstrated signature network assembly, functional analysis and logical operations on the networks that can be generated. In addition, we have used this technique in a proof of concept study to compare the effect of differential drug treatment using 4-hydroxytamoxifen and estrogen on the MCF-7 breast cancer cell line from a previously published study. We have shown that the two compounds can be differentiated by the networks of interacting genes. Both networks consist of a core module of genes including *c-Fos* as part of *c-Fos/c-Jun* heterodimer and *c-Myc* which is clearly visible. Using algorithms in our MetaCore™ software we are able to subtract the 4-hydroxytamoxifen and estrogen networks to further understand differences between these two treatments and show that the estrogen network is assembled around the core with other modules essential for all phases of the cell cycle. For example, Cyclin D1 is present in networks for the estrogen treated cells from two separate studies. These signature networks represent an approach to identify biomarkers and a general approach for discovering new relationships in complex high throughput toxicology data.

© 2005 Published by Elsevier Ireland Ltd.

Keywords: Signature networks; Microarray; Networks; Algorithms; Software

In recent years, the development of systems biology, databases and pathway software has enabled the analysis of high-throughput data in the context of the whole

cell. Microarray gene expression profiling is applied to sub-categorize complex diseases, predict treatment outcomes, and disease progression using “gene signatures” which represent differentially expressed genes that correlate with phenotype following clustering analysis or other statistical procedures (van't Veer et al., 2002; Chang et al., 2003). The assumption is that compounds associated with particular toxic end-points in

* Corresponding author. Tel.: +1 269 930 0974;
fax: +1 215 481 0159.

E-mail addresses: sean@genego.com, ekinssean@yahoo.com
(S. Ekins).

human would cause similar gene expression patterns in pre-clinically used mammals and in vitro systems. The human toxicity of a new compound can then be extrapolated by comparison to a database of expression profiles of compounds with known toxic end-points. Changes in global gene expression patterns in animal or cells at multiple dose levels and time-points produce relevant “signature” genes which have been used as predictors in humans (Hamadeh et al., 2002; Waters et al., 2003). This approach is powerful yet limited by its sensitivity to experimental design and statistical treatment (Huang et al., 2003).

The reproducibility and consistency of gene expression data has been demonstrated by different industrial and academic laboratories using commercial and custom arrays (Chu et al., 2004) and has been applied in hepatotoxicity, nephrotoxicity, cardiology and genotoxicity studies (Amin et al., 2004; Kramer et al., 2004), demonstrating the reliable identification of major cellular subsystems involved in toxic responses (Ulrich et al., 2004). Most biologically active compounds impact expression of multiple levels of cellular functionality, for example, including endogenous metabolic pathways, enzymes of phases I and II drug metabolism, nuclear hormone receptors (NHR) regulating cytochrome P450 enzymes (CYP), transporters and other genes (Ekins et al., 2005b). Gene expression patterns can also be used for elucidating new hypotheses for toxicity mechanisms (Gerhold et al., 2001; Heinloth et al., 2004; Lord, 2004).

The analysis of the existing databases on metabolic and signaling pathways shows that if taken together, known reactions and interactions would form a large cluster linked via molecular nodes shared among many process (Jeong et al., 2000). As direct dynamical modeling is currently limited to well characterized systems, a novel approach for making sense of large gene networks has been developed in recent years. This method combines graph theory and statistical methods for finding network modules that represent potential active functional units. One important result achieved in this area is the realization that biological networks of different origins (e.g., metabolic, regulatory, protein interactions, networks for different organisms) share the same global architecture (Giot et al., 2003; Barabasi and Oltvai, 2004). One consequence of such an architecture is the presence of so-called hub nodes which are high degree nodes connected to many low degree

nodes (Barabasi and Oltvai, 2004). Such topological features give networks the property of robustness with removal of even a substantial fraction of nodes leaving the network connected (Albert et al., 2000).

Our approach described herein represents an alternative microarray analysis method that connects interacting differentially expressed genes into condition-specific functional “signature networks” independent of statistical procedures and can yield the global cellular mechanisms behind differences in expression. This method is rooted in the idea of modular organization of large scale networks of biological processes proposed by Hartwell et al. (1999). According to this theory, various types of cellular functionality are provided by relatively small, transient but tightly connected networks of molecules (5–25 nodes) that are engaged in performing specific functions. We will demonstrate as a proof of concept study how a commercial software platform (MetaCore™), built around an integrated Oracle-based software platform can be used to define signature networks for published microarray data using 4-hydroxytamoxifen and estrogen treatments of the MCF-7 breast cancer cell line.

1. Materials and methods

1.1. Data annotation and software programming

An interactive, manually annotated database was derived from literature publications on proteins and small molecules (MetaCore™, GeneGo, St. Joseph, MI). This was developed with an Oracle version 9.2.0.4 Standard Edition (Oracle, Redwood Shores, CA) based architecture for the representation of biological functionality and integration of functional, molecular, or clinical information (Bugrim et al., 2004). Functional processes are the core objects in the database which can be of a different types and have unique relationships with molecular entities. The novel database architecture that is the foundation of MetaCore™ and a second product MetaDrug™ has been described in detail previously (Bugrim et al., 2004; Ekins et al., 2005a,b). To briefly summarize, functional processes and components serve as the core information space-holders in our database with many-to-many relationships between them. The corresponding molecular and mechanistic data are then linked to these space-holders as

they become available. Functions serve as the “linking portals” for heterogeneous data. Once linked, the heterogeneous types of high-throughput data become a part of a larger system-level picture where functional relations among them can be more easily established and elucidated (e.g., all proteins in a pathway and their genes with expression patterns). Every pathway and its elements (interactions, reactions, enzymatic functions) are linked to available molecular data (genes, proteins, compounds, expression data, SNPs, etc.) and annotated with relevant information about their involvement and importance in a number of common human diseases.

The MetaCore™ software runs on an Intel-based 32 bit server running RedHat Linux Enterprise 3 AS (RedHat, Raleigh, NC) and the web server ran Apache 1.3.x/mod_perl (<http://perl.apache.org/start/index.html>). Software on the server side was written in Perl while the client side required HTML/JavaScript and the Macromedia Flash Player Plug-in (Macromedia Inc., San Francisco, CA).

1.2. Visualization of microarray data for 4-hydroxytamoxifen and estrogen treated breast cancer cell line on gene networks

We have used data from a published study of G₀-arrested MCF-7 breast cancer cells treated with estrogen (E₂) and 4-hydroxytamoxifen (OHT) using the NIEHS ToxChip microarray consisting of 1901 genes (Hodges et al., 2003). This file was obtained from the NIEHS website <http://dir.niehs.nih.gov/microarray/datasets/home-pub.htm>. The microarray data describing the up/down regulation of genes at 0.5, 2, 4, 12 and 24 h time points (Hodges et al., 2003) was imported into MetaCore™ on the client side as a tab-delimited file and genes showing statistically significant expression changes with treatments were selected. Several algorithms to enable both the construction and analysis of gene networks are integrated within MetaCore™ and have been previously described for MetaDrug™ (Ekins et al., 2005a–c). In this study, the proteins were queried against the MetaCore™ database and networks were constructed for time points using a strict “direct interactions” algorithm (Only the selected genes are included in the networks; in this case those over- and under-expressed with certain thresholds). The individual networks had discrete gene content (not shown) and were combined

into integrated networks for each treatment, revealing all possible gene connections affected over 24 h ($p < 0.0005$). Genes from the microarray that were up or down regulated could then be visualized on this network.

A second dataset for E₂ treatment of MCF-7 cells for 1, 4, 12, 24, 36 and 48 h (Lobenhofer et al., 2002), with gene expression changes assessed using the same ToxChip was used. In this case 105 genes were shown to be estrogen regulated at all six time points. We have used these genes as a starting point for generating a combined network. The published gene list (Lobenhofer et al., 2002) was first converted to GenBank accession numbers from the NIEHS website (<https://dir-apps.niehs.nih.gov/maps/guest/clonesrch.cfm>), then converted to Locuslink identifiers using DAVID (Dennis et al., 2003) (<http://apps1.niaid.nih.gov/david/>). The expression intensity values were set as +1 or –1 for up and down regulation, respectively. The tab delimited file was then uploaded in MetaCore™ to generate a gene network with the “direct interactions” algorithm.

2. Results

The original 2-D hierarchical clustering at 0.5, 2, 4, 12 and 24 h time points showed both compounds caused nearly identical expression patterns at early time points, featuring induction of cell cycle progression genes (Hodges et al., 2003) (Fig. 1A). Ultimately, E₂ treatment leads to cell proliferation, while OHT treated cells largely remain arrested at G₀. An additional supervised analysis (applicable when the distinct phenotypes are well defined (Quackenbush, 2001)) revealed E₂ induced expression of Cyclin D1, a key initiator of G₀/G₁ and G₂/M progression (Stacey, 2003) and other genes escaping the original clustering. We hypothesized that network analysis of raw microarray data would be more informative as this would include key regulatory genes that would otherwise be removed during clustering, in turn resulting in unique gene-network topologies. After the microarray data (Hodges et al., 2003) was imported into MetaCore™ the integrated networks were strikingly different but both contained early transcriptional factors *c-Myc*, *c-Jun* and *c-Fos*. Despite using a small array excluding many cell cycle genes, the E₂ network (Fig. 1B) featured induced genes essential for all cell

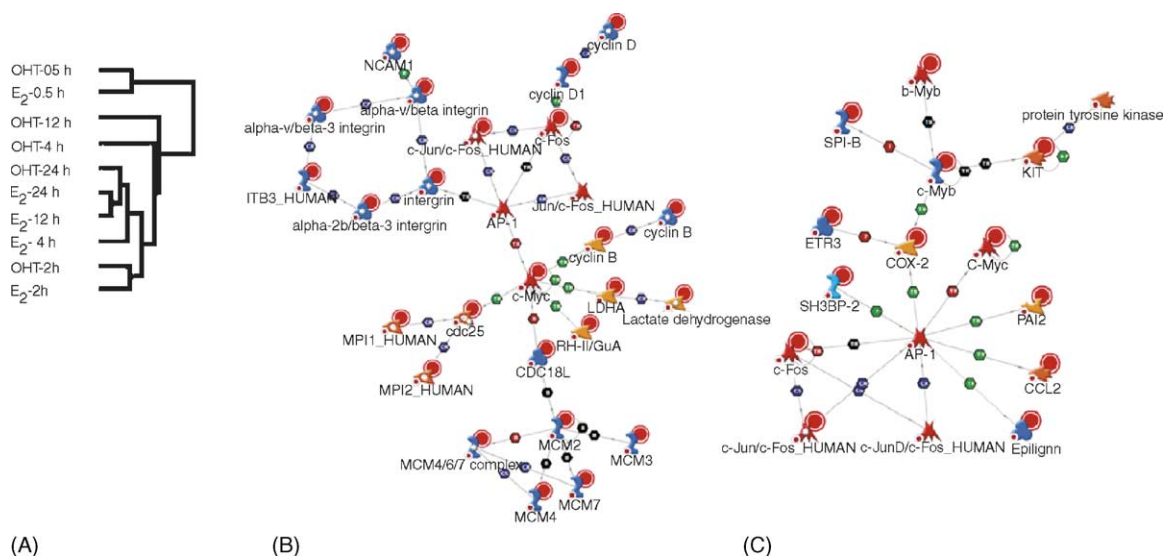


Fig. 1. (A) Published hierarchical clustering of overexpressed genes show virtually no difference between treatments (Hodges et al., 2003). (B) Gene network signatures after applying the direct interactions algorithm to the list of overexpressed genes obtained from microarrays of MCF-7 breast cancer cells following treatment with estrogen and (C) 4-hydroxytamoxifen (B). Colored symbols represent genes (network nodes) and red solid circles represent genes with microarray data showing significant upregulation. Small colored hexagons on vectors between nodes describe positive (green), negative (red), unspecified (black), interactions or logical relationships (blue).

cycle phases including Cyclin D1 (G_0/G_1 and G_2/M transition), subunits of the DNA replication licensing complex MCM (S phase), cyclin B1 (G_2/M) and *cdc25*-encoding genes (M phase inducers). None of these proteins were present on the OHT network (Fig. 1C). The differences in the “signature networks” correlated with distinct cellular responses to treatments, without previous knowledge of the end-points over 24 h. This analysis enables up- and down-regulated genes to be visualized on the same network identifying key “switcher” nodes (where, the “sign” of expression changes from induction to repression) and potential therapeutic targets. Secondly, genes with sub-threshold or unchanged expression levels can be included as they may be central to pathways unaffected at the gene expression level, but result in downstream induction of genes. Logical operations can also be performed on networks using several algorithms in MetaCore™ to demonstrate that the intersection between the E₂ and OHT networks consists of *c-Fos* as part of *c-Fos/c-Jun* heterodimer and *c-Myc* (Fig. 2). This represents a core module shared by both compound treatments. In addition, the OHT network can be subtracted from the estrogen network (Fig. 3A) or the estrogen network can be subtracted

from the OHT network (Fig. 3B). In summary, the estrogen network is assembled around the core of *c-Fos/c-Jun* and *c-Myc* while other integrated modules include proteins essential for all phases of the cell cycle which are highlighted individually (Fig. 4).

A second dataset for E₂ treatment of MCF-7 cells for 1, 4, 12, 24, 36 and 48 h (Lobenhofer et al., 2002) consisting of 105 genes that had been shown to be regulated by estrogen at all six time points was further used to evaluate MetaCore™. Out of the initial 105 genes, only 94 were found with LocusLink identifiers and of these 3 did not map to our database (Locuslink ID 50485, 5203, 9231). The remaining genes that were successfully mapped to our database were used to generate a combined network with the “direct interactions” algorithm, resulting in two relatively large networks with connected genes (Fig. 5). One of these networks contained genes related to apoptosis such as APP and Cathepsin D (Fig. 5A) while a second network contained several of the same genes previously identified on the estrogen signature network, namely those involved primarily in the cell cycle Cyclin D1, Cyclin D, *c-Myc*, LDHA and CDC18L (Fig. 5B).

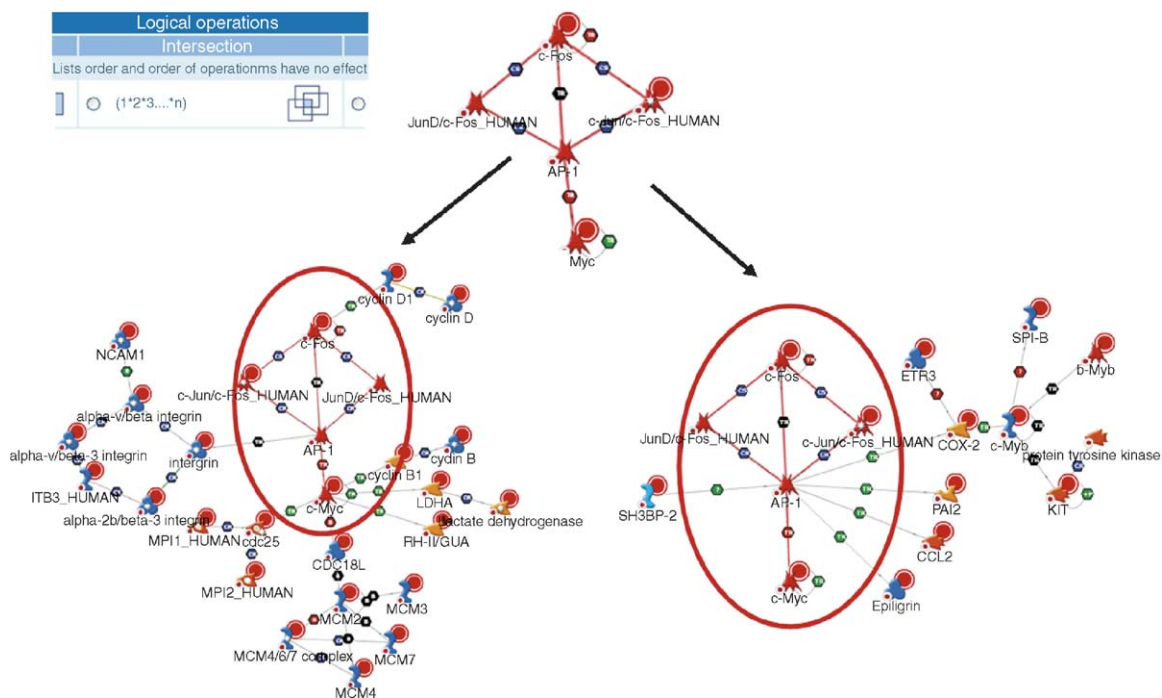


Fig. 2. Intersection between the estrogen and tamoxifen networks (described graphically in the box) consists of *c-Fos* as part of *c-Fos/c-Jun* heterodimer and *c-Myc*. Colored symbols represent genes (network nodes) and red solid circles represent genes with microarray data showing significant upregulation. Small colored hexagons on vectors between nodes describe interactions. A core module is highlighted in red on both the estrogen and 4-hydroxytamoxifen networks. Small colored hexagons on vectors between nodes describe positive (green), negative (red), unspecified (black) interactions or logical relationships (blue).

3. Discussion

Traditionally, the functional organization of a biological system has been described in terms of pathways. Since Hartwell et al. (1999) hypothesized that many biological functions are carried out by discrete modules of physically interacting proteins (Hartwell et al., 1999), searching for interacting proteins and networks has become a key use for the functional analysis of microarrays. It is well documented that the outcome of microarray analysis is largely dependent upon statistical procedures applied both at the level of determining sets of genes with significantly altered expression and, to a greater extent, at the level of data clustering (Butte, 2002). In general, different combinations of several standard statistical procedures have been used for the analysis of gene expression patterns on microarrays: hierarchical clustering (Eisen et al., 1998), *k*-value clustering (Tavazoie et al., 1999) self-organizing maps (SOMs) (Tamayo et al., 1999), principal component

analysis (PCA) and analysis of variance (ANOVA). Eventually, the expression pattern is represented as a tree with branch length corresponding to the degree of similarity between conditional expression. The analysis identifies those sets of genes for which expression is the most distinct between the classes and may result in the “signature sets” of genes for certain phenotypes as determined by such supervised procedures.

While understanding the global gene network architecture is an important first step, from the practical standpoint of toxicogenomics, it is even more interesting to find the specific functional modules that are implemented in toxicity mechanisms. A recent review has described the theory behind biological network models and their architecture (Barabasi and Oltvai, 2004). Some algorithms can link the expressed genes with known transcriptional factors via the genome-wide identification of DNA binding sites (Bar-Joseph et al., 2003) or enable building of condition-specific probabilistic models (Segal et al., 2003). The

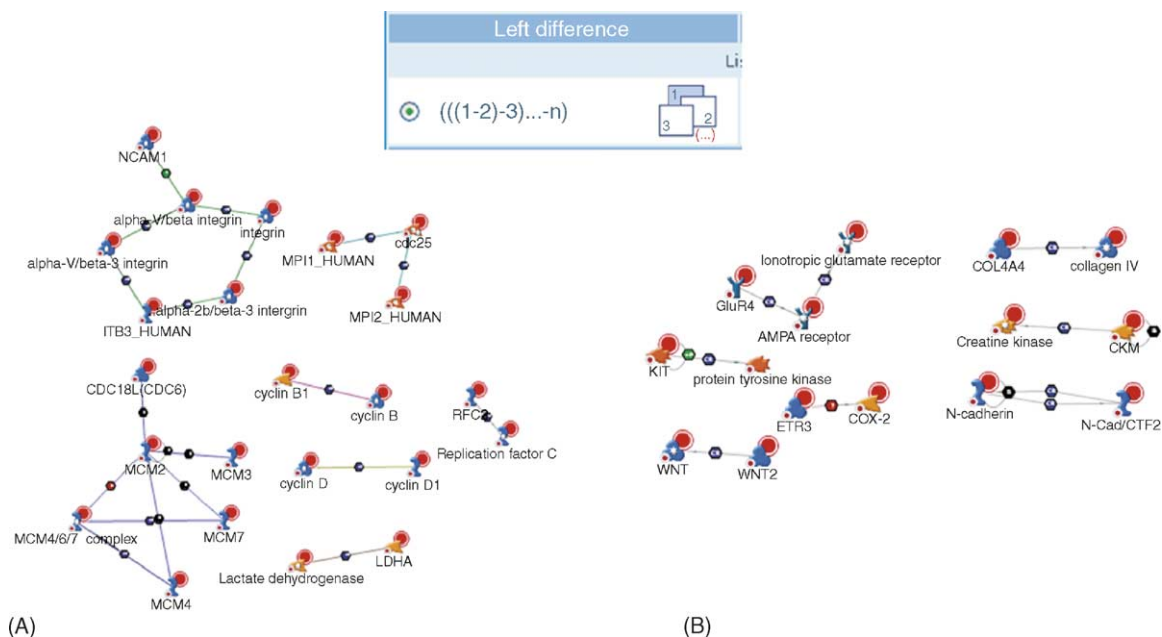


Fig. 3. (A) Subtraction of the tamoxifen network from the estrogen network (described graphically in the box) and (B) subtraction of the estrogen network from the tamoxifen network. Colored symbols represent genes (network nodes) and red solid circles represent genes with microarray data showing significant upregulation. Small colored hexagons on vectors between nodes describe positive (green), negative (red), unspecified (black) interactions or logical relationships (blue).

co-expressed genes can be grouped based on the occurrence of common *cis*-elements in the context of reference databases of transcriptional factors (Matys et al., 2003) of well-known metabolic pathways (Ihmels et al., 2002). Microarray data is, therefore often linked to curated knowledge databases on proteins, genes and pathways such as SwissProt, KEGG, MIPS and others. The genes and proteins can be grouped and networked together based on their names or co-occurrence in MedLine texts using Natural Language Processing, NLP (Jenssen et al., 2001; Chaussabel and Sher, 2002). However, only 30%–40% of experimentally verified protein interactions were reported to correspond to pairs from MedLine abstracts (Blaschke and Valencia, 2001). There are more biologically relevant sources of information on interacting proteins such as curated databases BIND (Biomolecular Interaction Network Database (Bader et al., 2004)), INTERACT (Elbeck et al., 1999), MIPS (Munich Information Center for Protein Sequences) (Mewes et al., 2002). The database of Interacting Proteins, DIP (Xenarios et al., 2000) and the Human protein references database (HPRD; Peri et al., 2003) which generally have rel-

atively low coverage, particularly for human-specific interactions. Several companies offer larger curated interaction databases covering human metabolism, regulation and signaling. Manually curated databases with interactions extracted from published experimental data on individual reactions are considered as the gold standard for benchmarking purposes (von Mering et al., 2002). Combining these comprehensive databases with powerful analytical and network building tools has resulted in the development of several integrated high-throughput data mining suites like Pathway AssistTM, MetaCoreTM, Pathways AnalysisTM and PathArtTM available from several companies. These tools can enable visualization of global cellular mechanisms behind differences in gene expression.

In this study, we have used one such suite, MetacoreTM which enables the user to: (1) upload and simultaneously visualize on the pathways microarray data of multiple formats including proteomics and metabolomics data; (2) generate networks using multiple algorithms based on the aforementioned types of datasets and a comprehensive curated database of human protein-protein and protein-compound interac-

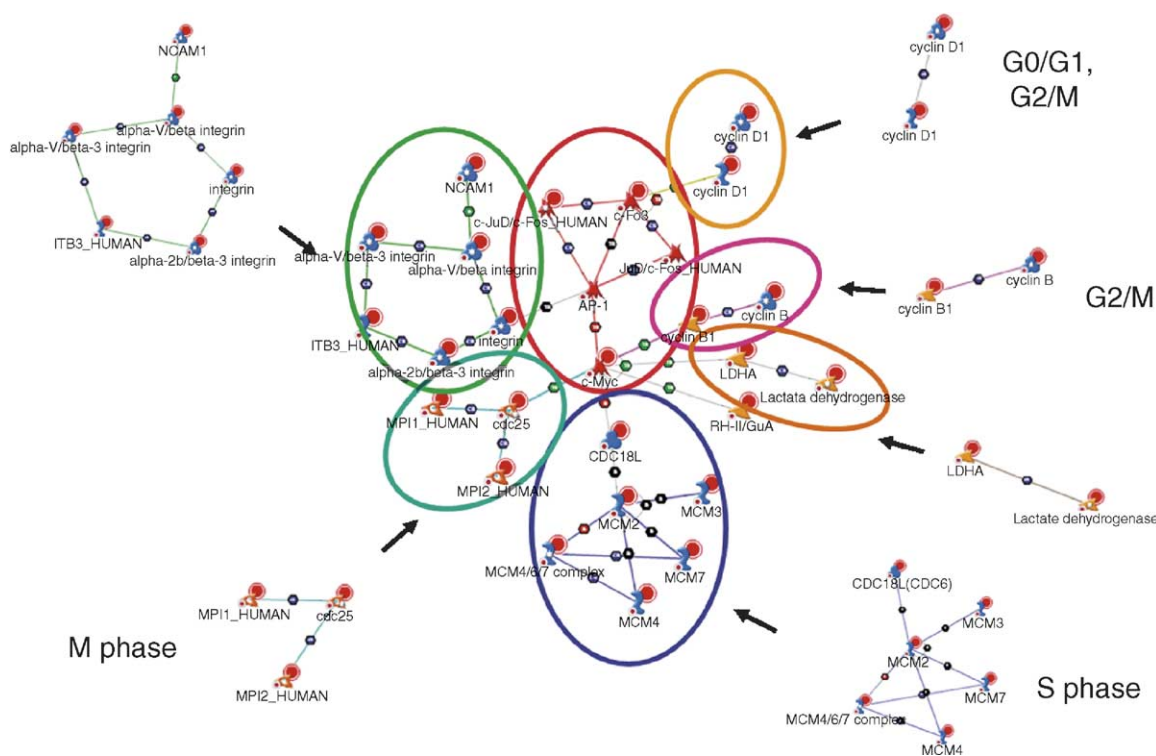


Fig. 4. The estrogen combined network is assembled around the core of *c-Fos/c-Jun* and *c-Myc*. The integrated modules include proteins essential for all phases of the cell cycle highlighted and shown separately. Colored symbols represent genes (network nodes) and red solid circles represent genes with microarray data showing significant upregulation. Small colored hexagons on vectors between nodes describe positive (green), negative (red), unspecified (black) interactions or logical relationships (blue).

tions database; (3) export/import data exchange with different informatics systems and seamlessly access the major public bioinformatics resources. The analysis in this software, therefore establishes the universe of direct physical interactions between active proteins in human cells using information extracted from the experimental literature. Additionally, this set of interactions defines the cellular potential for assembly of protein complexes, signaling and effector pathways. By positioning expression data onto experimentally verified physical interactions, we effectively restrict all possible associations between array spots to those that are physically possible in the cell. This then produces the condition-specific functional “signature networks” which are sets of functional pathways organized into networks of a certain topology, which are activated (individual genes on the pathways can be up- or down-regulated) in a specific condition.

The types of network analysis described in this study using MetaCore™, therefore do not require data clustering. We have illustrated the application of this approach using a published study with raw microarray data from MCF-7 cells treated with E₂ and OHT at different time points. Although the authors of this study had used hierarchical clustering to show the similarity of gene expression patterns for both treatments at the earliest time points, there is divergence later due to the known effects of E₂ on cell proliferation, while OHT treated cells largely remain arrested at G₀. The present proof of concept study was undertaken to show how different compound treatments could be illustrated and has shown that both can be differentiated by the networks of interacting genes. In particular the OHT network includes several genes that were not upregulated (Fig. 1C). In addition, both networks consist of a core module of genes including *c-Fos* as part of *c-Fos/c-*

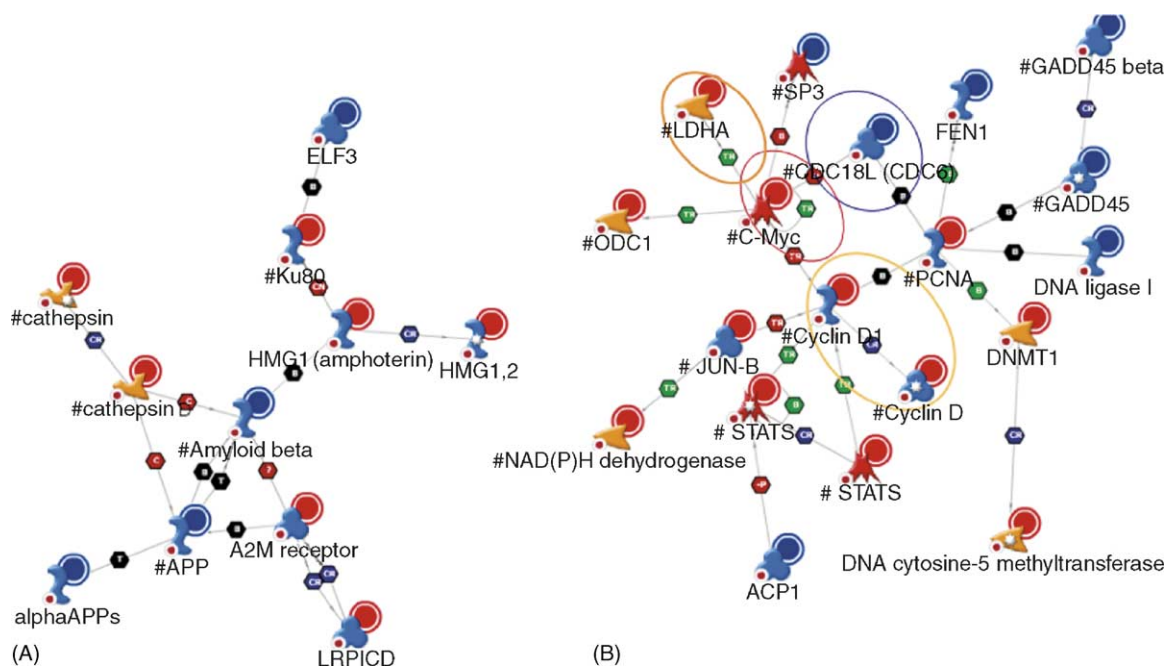


Fig. 5. The most connected networks resulting from analysis of genes shown to be estrogen responsive after treatment of MCF-7 cells for 1, 4, 12, 24, 36 and 48 h (Lobenhofer et al., 2002). Analysis within MetaCore™ resulted in two large networks with connected up and down regulated genes (A, B). The highlighted regions on network B correspond to the genes present in the sections for the combined estrogen network in Fig. 4, in particular the central gene Cyclin D1 and others related to phases of the cell cycle. Colored symbols represent genes (network nodes) and red solid circles represent genes with microarray data showing significant upregulation while blue solid circles show significant downregulation. Small colored hexagons on vectors between nodes describe positive (green), negative (red), unspecified (black) interactions or logical relationships (blue).

Jun heterodimer and *c-Myc* (Fig. 2). Previously it has been shown that E_2 regulates *c-Myc* and Cyclin D1 to promote G_1/S transition (Foster et al., 2001) while the importance of increased Cyclin D1 expression as a rate limiting step in the cell cycle has been widely described (Coqueret, 2002; Lobenhofer et al., 2002; Stacey, 2003). It is perhaps, not surprising that there are common transcriptional events between these two compounds as OHT is a mixed agonist. Using algorithms in the software we are able to subtract networks to further understand differences and show that the E_2 network is assembled around the core module described above with other modules essential for all phases of the cell cycle (Fig. 4). These signature networks represent a method to identify biomarkers which provides a general approach for discovering new relationships in complex data. Secondly, compounds showing comparable effects on gene expression as either of these two compounds would be expected to produce similar

signature networks as either E_2 or OHT after using the methods described.

A second dataset for E_2 treatment of MCF-7 cells was also obtained from the literature (Lobenhofer et al., 2002). In this case the treatments were for 1, 4, 12, 24, 36 and 48 h and instead of using the whole set of 1901 genes, we used the subset of 105 genes suggested by the authors to have significant induction or repression in order to generate a signature network for E_2 . Interestingly, two discrete networks were indicated, one of which had some of the genes previously observed in the estrogen signature network from a separate study, including the important Cyclin D1 and other cell cycle genes (Fig. 5B). The differences between the estrogen networks generated for each dataset are likely related to either the statistical filtering of the data, inclusion of induced and repressed genes or the longer estrogen exposure times applied in the second study. A more exhaustive analysis of the raw microarray dataset

for this study and selection of comparable time points only (excluding 36 and 48 h data) as performed with the E₂ and OHT study, may result in more comparable signature networks for estrogen between the two studies.

In summary, we have described an approach to signature network construction that does not have any of the limitations of statistical approaches to identifying signature genes. The proof of principle study would indicate that this software based on a manually curated database of interactions and network building algorithms can be more widely applied in toxicogenomics for the analysis of high throughput data.

Acknowledgements

Dr. Craig Giroux (Wayne State University) and Dr. Christopher Bradfield (McCardle Cancer Institute) are acknowledged for their support and discussions. Our colleagues at GeneGo, Drs. Sergey Andreyev, Svetlana Sorokina, Tatyana Serebrijskaya, Roman Zuev, Andrej Ryabov, Eugene Kirillov, Eugene A. Rakhmatulin are sincerely acknowledged for their continuing contributions to software development and data annotation.

References

- Albert, R., Jeong, H., Barabasi, A.L., 2000. Error and attack tolerance of complex networks. *Nature* 406, 378–382.
- Amin, R.P., Vickers, A.E., Sistare, F., Thompson, K.L., Roman, R.J., Lawton, M., Kramer, J., Hamadeh, H.K., Collins, J., Grissom, S., Bennett, L., Tucker, C.J., Wild, S., Kind, C., Oreffo, V., Davis 2nd, J.W., Curtiss, S., Naciff, J.M., Cunningham, M., Tennant, R., Stevens, J., Car, B., Bertram, T.A., Afshari, C.A., 2004. Identification of putative gene based markers of renal toxicity. *Environ. Health Perspect.* 112, 465–479.
- Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J., 2004. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85.
- Barabasi, A.-L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113.
- Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K., Jaakola, T.S., 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10146–10151.
- Blaschke, K., Valencia, A., 2001. Can bibliographical pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Funct. Genomics* 2, 196–206.
- Bugrim, A., Nikolskaya, T., Nikolsky, Y., 2004. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov. Today* 9, 127–135.
- Butte, A., 2002. The use and analysis of microarray data. *Nat. Rev. Drug Discov.* 1, 951–960.
- Chang, J.C., Wooten, E.C., Tsimelzon, A., Hilsenbeck, S.G., 2003. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 362, 362–369.
- Chaussabel, D., Sher, A., 2002. Mining microarray expression data by literature profiling. *Genome Biol.* 3, 0055.0051-0055.0016.
- Chu, T.M., Deng, S., Wolfinger, R., Paules, R.S., Hamadeh, H.K., 2004. Cross-site comparison of gene expression data reveals high similarity. *Environ. Health Perspect.* 112, 449–455.
- Coqueret, O., 2002. Linking cyclins to transcriptional control. *Gene* 299, 35–55.
- Dennis Jr., G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* 4, 3.
- Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868.
- Ekins, S., Bugrim, A., Nikolsky, Y., Nikolskaya, T., 2005a. Systems biology: applications in drug discovery. In: Gad, S. (Ed.), *Drug Discovery Handbook*. Wiley, New York, pp. 123–183 (Chapter 4).
- Ekins, S., Kirillov, E., Rakhmatulin, E.A., Nikolskaya, T., 2005b. A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism. *Drug Metab. Dispos.* 33, 474–481.
- Ekins, S., Nikolsky, Y., Nikolskaya, T., 2005c. Techniques: application of systems biology to absorption, distribution, metabolism, excretion, and toxicity. *Trends Pharmacol. Sci.* 26, 202–209.
- Elbeck, K., Brass, A., Paton, N., Hodgman, C., 1999. INTERACT: an object-oriented protein-protein interactions database. *Proc. Int. Conf. Intel. Syst. Mol. Biol.* 87–94.
- Foster, J.S., Henley, D.C., Ahamed, S., Wimalasena, J., 2001. Estrogens and cell-cycle regulation in breast cancer. *Trends Endocrinol. Metab.* 12, 320–327.
- Gerhold, D., Lu, M., Xu, J., Austin, C., Caskey, C.T., Rushmore, T., 2001. Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. *Physiol. Genomics* 5, 161–170.
- Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrola, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley Jr., R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M., 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736.
- Hamadeh, H.K., Bushel, P.R., Jayadev, S., DiSorbo, O., Bennett, L., Li, L., Tennant, R., Stoll, R., Barrett, J.C., Paules, R.S., Blanchard, K., Afshari, C.A., 2002. Prediction of compound signa-

- ture using high density gene expression profiling. *Toxicol. Sci.* 67, 232–240.
- Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., 1999. From molecular to modular cell biology. *Nature* 402, C47–C52.
- Heinloth, A.N., Irwin, R.D., Boorman, G.A., Nettesheim, P., Fannin, R.D., Sieber, S.O., Snell, M.L., Tucker, C.J., Li, L., Travlos, G.S., Vansant, G., Blackshear, P.E., Tennant, R.W., Cunningham, M.L., Paules, R.S., 2004. Gene expression profiling of rat livers reveals indicators of potential adverse effects. *Toxicol. Sci.* 80, 193–202.
- Hodges, L.C., Cook, J.D., Lobenhofer, E.K., Li, L., Bennett, L., Bushel, P.R., Aldaz, C.M., Afshari, C.A., Walker, C.L., 2003. Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells. *Mol. Cancer Res.* 1, 300–311.
- Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., West, M., Nevins, J.R., Huan, A.T., 2003. Gene expression predictors of breast cancer outcomes. *Lancet* 361, 1590–1596.
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N., 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31, 370–377.
- Jenssen, T.K., Laegreid, A., Komorowski, J., Hovog, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L., 2000. The large-scale organization of metabolic networks. *Nature* 407, 651–654.
- Kramer, J.A., Pettit, S.D., Amin, R.P., Bertram, T.A., Car, B., Cunningham, M., Curtiss, S.W., Davis, J.W., Kind, C., Lawton, M., Naciff, J.M., Oreffo, V., Roman, R.J., Sistare, F.D., Stevens, J., Thompson, K., Vickers, A.E., Wild, S., Afshari, C.A., 2004. Overview on the application of transcription profiling using selected nephrotoxicants for toxicology assessment. *Environ. Health Perspect.* 112, 460–464.
- Lobenhofer, E.K., Bennett, L., Cable, P.L., Li, L., Bushel, P.R., Afshari, C.A., 2002. Regulation of DNA replication fork genes by 17beta-estradiol. *Mol. Endocrinol.* 16, 1215–1229.
- Lord, P.G., 2004. Progress in applying genomics in drug development. *Toxicol. Lett.* 149, 371–375.
- Matys, V., Fricke, E., Geffers, R., Goblting, E., Haubrock, M., Hehl, R., 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *NAR* 1, pp. 374–378.
- Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., Weil, B., 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34.
- Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobel, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., Pandey, A., 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13, 2363–2371.
- Quackenbush, J., 2001. Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427.
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Freidman, N., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176.
- Stacey, D.W., 2003. Cyclin D1 serves as a cell cycle regulatory switch in actively proliferating cells. *Curr. Opin. Cell Biol.* 15, 158–163.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2907–2912.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–286.
- Ulrich, R.G., Rockett, J.C., Gibson, G.G., Pettit, S.D., 2004. Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. *Environ. Health Perspect.* 112, 423–427.
- van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P., 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403.
- Waters, M., Boorman, G., Bushel, P., Cunningham, M., Irwin, R., Merrick, A., Olden, K., Paules, R., Selkirk, J., Stasiewicz, S., Weis, B., Van Houten, B., Walker, N., Tennant, R., 2003. Systems toxicology and the chemical effects in biological systems (CEBS) knowledge base. *EHP Toxicogenomics* 111, 15–28.
- Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Markotte, E.M., Eisenberg, D., 2000. DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291.