34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Revealing the dynamic modularity of composite biological networks in breast cancer treatment*

Konstantina Dimitrakopoulou, *Student Member*, George Dimitrakopoulos, *Student Member*, Evangelia I. Zacharaki, Ioannis A. Maraziotis, Kyriakos Sgarbas, *Member,* and Anastasios Bezerianos, *SeniorMember, IEEE*

Abstract—**A major challenge in modern breast cancer treatment is to unravel the effect of drug activity through the systematic rewiring of cellular networks over time. Here, we illustrate the efficacy and discriminative power of our integrative approach in detecting modules that represent the regulatory effect of tamoxifen, widely used in anti-estrogen treatment, on transcriptome and proteome and serve as dynamic sub-network signatures. Initially, composite networks, after integrating protein interaction and time series gene expression data between two conditions (estradiol and estradiol plus tamoxifen), were constructed. Further, the Detect Module from Seed Protein (DMSP) algorithm elaborated on the graphs and constructed modules, with specific 'seed' proteins used as starting points. Our findings provide evidence about the way drugs perturb and rewire the high-order organization of interactome in time.**

I. Introduction

Recent systems biology studies have shifted their interest to rationalize complex diseases, from analyzing individual biological components to networks of molecules. Accumulating evidence suggests that alterations in the apparent modularity (i.e. the existence of interacting, separable and functional groups of genes/proteins) governing these networks enable the identification of sub-network disease biomarkers. These graph-theoretic approaches assisted significantly in the comprehension of cancer pathogenesis, progression and metastasis [1], establishing thus the necessity of the Systems Biology field in clinical practice (Systems Medicine). The ultimate goal of Systems Medicine is to provide diagnostic and prognostic biomarkers, identify disease subtypes and set the optimized treatment, leading thus to a better and more personalized medicine.

The motive of this study is multifarious; the main intrigue lies in the fact that the majority of breast cancer patients that express estrogen-receptor alpha (ERα) usually undergo tamoxifen treatment, with no good outcome in all cases; also, obscure remains the scene where similar gene expression patterns among patients with regard to known gene markers cannot guarantee similar phenotype (i.e. disease outcome). Late studies implicated that alterations in gene expression might perturb the higher-level organization of the interactome, affecting so the disease outcome [2]. To investigate this hypothesis, we explored how the temporal dynamics of transcriptional behavior in a specific treatment scheme (estradiol and estradiol plus tamoxifen) reforms the protein interactome. Our target was to reveal how interactome 'areas', in the form of modules/sub-networks, are perturbed in response to drug over time. To achieve this goal, we integrated gene expression and protein-protein interaction (PPI) data, a strategy recently established as fruitful in providing information of specific genes/proteins on disease-specific pathophysiology. Towards this orientation, many studies have combined multiple data types [3, 4]; works like [5] scored pathways based on the similarity of the expression values of the participating pathway genes. For example, interesting studies like [6] detected sub-networks of highly co-expressed genes on the protein graph by starting from a random gene with the use of a greedy algorithm, which cannot guarantee completeness. Other studies like [7] integrated gene expression, PPI and phenotype data to identify dense modules with the provision of incorporating additional constraints from a variety of datasets. However, this approach is primarily designed for finding protein complexes from protein interaction data, is sensitive to gene expression noise and promotes the detection of dense modules.

In this paper, we illustrate the efficacy of our integrative methodology [8] in capturing the dynamic modular transitions in response to tamoxifen. To achieve this goal, we reinforced the protein graph structure, via weighting scheme, with time series microarray data descending from an in vivo study [9]. Next, the Detect Module from Seed Protein (DMSP) algorithm defined modules on the composite protein network starting from specific 'seed' proteins. An important feature of this algorithm is that the overlaid gene expression information, in the form of weight, reassures the entrance of certain interactions into the modules, even if they are not favored by the topology. Also, DMSP saves many interactions among proteins that interact closely (e.g. complexes) even if they show dissimilar or inverse expression trends, through the rest weighted neighbors of such an interaction.

Our time-evolving modules report that the response to tamoxifen is a highly dynamic process and raise several biological questions regarding the recruitment of several known pathways. Finally, our findings corroborate towards the integration of heterogeneous data and the detection of discriminative temporal sub-networks that serve as hallmarks of disease-specific states.

*Research co-financed by the European Union (European Social Fund–ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

K.D., E.I.Z., I.A.M. and A.B. are with the Medical School, Patras, 26500 GR (corresponding author phone: +30-2610-969147; e-mail: kondim@upatras.gr, ezachar@upatras.gr, i.maraziotis@gmail.com, bezer@upatras.gr).

G.D and K.S. are with the Department of Electrical and Electronic Engineering, Patras, 26500 GR (e-mail: geodimitrak@upatras.gr, sgarbas@upatras.gr).

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

II. Methods

A. Data sets and pre-processing

We downloaded all human protein interaction data from HPRD (http://www.hprd.org/), BioGRID (http://thebiogrid.org/), IntAct (http://www.ebi.ac.uk/intact/) and InnateDB (http://www.innatedb.ca/) databases. Regarding gene expression data we used the raw time series microarray data (days 1, 2, 4, 7, 14) publicly available in NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE22386 [9]. In particular, we isolated for analysis only the probe sets of genes that are present in the protein interaction network and ended up with 3307 gene names and 14498 interactions among them. The datasets (estradiol and estradiol plus tamoxifen treatment) were normalized after background correction with loess normalization approach with the use of limma package in Bioconductor [10]. The expression value of each gene was computed by taking the average of the corresponding probe sets and all values were normalized with respect to day1.

Finally, we downloaded all proteins related to breast cancer from G2SBC (http://www.itb.cnr.it/breastcancer/) and dbDEPC (http://lifecenter.sgst.cn/dbdepc/index.do) databases and 883 proteins were mapped to our final gene list. This subset defined the 'seed' list that was used as input to the DMSP algorithm.

B. Weighted Graph

The initial step of the weighting scheme includes clustering of the expression profiles of both datasets. In detail, we clustered the temporal profiles with the use of k-means algorithm that was able to process fast and transparently the datasets. The clustering process was repeated more than 100 times using random initialization, with Euclidean metric as distance measure. The number of clusters was appointed at 28 clusters with the use of Dunn index.

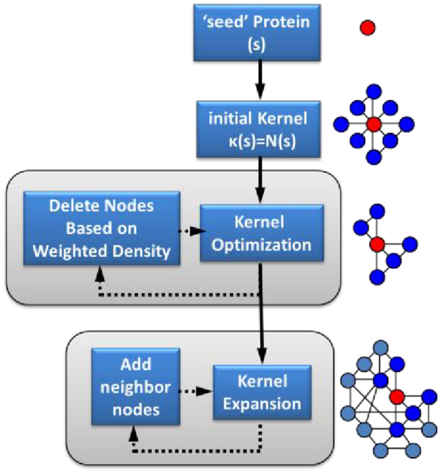

Figure 1. Outline of DMSP algorithm in constructing a module.

The human protein interaction graph is represented as a graph G(V,E). The vertices are the set of unique proteins (in our case |V|= 3307) and the edges are the interactions among the vertices (|E|=14498). In order to add weight to an interaction between two proteins x and y, we find the clusters C(x) and C(y) where they belong and the corresponding centroids K_x, K_y of these clusters. Then, we calculate the distance of each gene from its centroid and the distance between the two centroids. The weight of the PPI interaction is given by the metric:

$$W(x,y)=n_1(\|x-K_x\|^2+\|y-K_y\|^2)+n_2\|K_x-K_y\|^2$$

||.|| represents the distance metric (in our case Euclidean). The constants n_1 and n_2 add an extra confidence score to the factors of the weight function. Driven by the fact that there is noise (outliers) in the gene expression profiles, we set n_1=0.3 and n_2= 0.7 in order to enforce the distance between centroids comparing to the distance of each gene from its centroid.

C. DMSP Algorithm

The algorithm operates in two phases. Initially, given a 'seed' protein, it selects a subset of its most promising first order neighbors, subsequently expands this initial kernel to

TABLE1 KEGG Pathway analysis of characteristic module pairs with fold change in size ≥ 2.

Seed Protein	E_2 Size	KEGG Pathway term (pathway /P-value)	E_2+Tam Size	KEGG Pathway term (pathway /P-value)
ERBB2	159	ErbB signaling pathway 4.2E-17; T cell receptor signaling pathway 6.4E-14; Fc gamma R-mediated phagocytosis 6.6E-14; Jak-STAT signaling pathway 9.5E-13	76	ErbB signaling pathway 4.1E-16; T cell receptor signaling pathway 3.2E-13; Natural killer cell mediated cytotoxicity 4.5E-13
LYN	11	B cell receptor signaling pathway 2.4E-4; Fc gamma R-mediated phagocytosis 4.9E-4	139	ErbB signaling pathway 6.7E-17;Jak-STAT signaling pathway 7.8E-14; T cell receptor signaling pathway 8.7E-14; Natural killer cell mediated cytotoxicity 4.0E-13; Fc gamma R-mediated phagocytosis 1.9E-11; B cell receptor signaling pathway 9.3E-11
LCK	10	–	131	T cell receptor signaling pathway 3.2E-14;ErbB signaling pathway 1.4E-13; Natural killer cell mediated cytotoxicity 1.7E-11; B cell receptor signaling pathway 6.3E-10

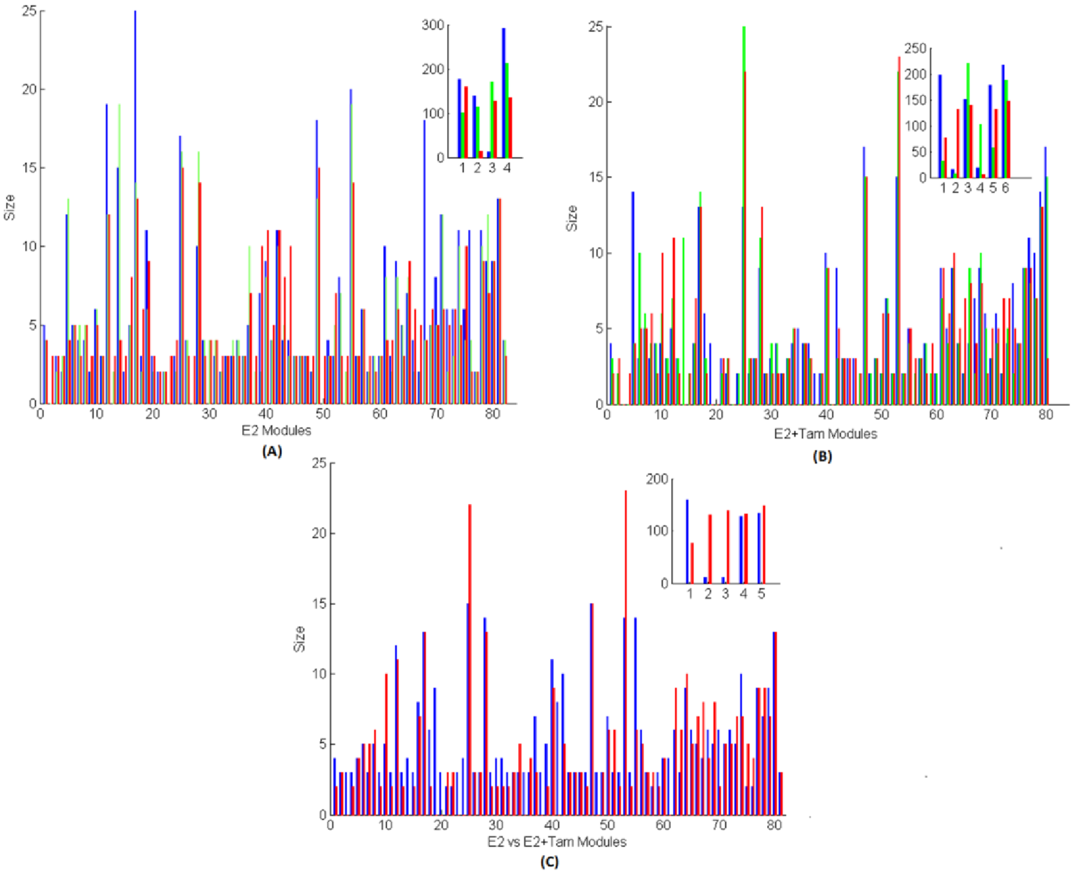

Figure 2. Bar plots displaying the size of modules descending from 86 seed proteins (the inset bar plots show the oversized module cases). (A) The size of E$_2$ modules in 3 time intervals (day_2–day_4: blue, day_2–day_7: green, day_2–day_14: red), (B) the size of E$_2$ plus tamoxifen modules in the aforementioned 3 time intervals. In (C) we compare the size of E$_2$ (blue) and E$_2$ plus tamoxifen modules (red) only in day_2–day_14 interval.

accept more proteins. This expansion is based on assumptions that take into account the internal, external, weighted internal and weighted external degrees of the nodes, concerning the number of neighbors for the specific protein as well as the weights of these connections (Fig.1). A detailed description is provided in [8]. The parameters of DMSP were set after exhaustive trials as $p_1 = 0.15$ and $p_2 = 0.5$.

Despite the fact that the weight metric favors interactions between genes with similar expression trends (i.e. small weight values indicate small distance among profiles) and thus promotes them in the topology, the DMSP algorithm manages to 'save' known interactions even if their expression profiles are inverse by incorporating information given by the rest weighted neighbors of such interactions, during the module construction.

III. RESULTS

Initially, after weighting the protein interaction networks with the gene expression values from all time points and applying the DMSP algorithm with 883 'seeds' as starting points, we ended up with 97 modules (i.e. only 97 seed proteins led to module construction with at least two members). Nevertheless, we isolated 86 modules with more than 3 members (at least in one type of treatment) for further analysis.

On second level, we zoomed into these 86 modules and run DMSP after weighting the PPI graphs with the gene expression values of three time intervals (day_2–day_4, day_2–day_7 and day_2–day_14) in order to detect the dynamic modular transitions every time a consecutive time point is added. Given the fact that 17β-estradiol (E$_2$) stimulates the growth of ER positive tumors, we expected significant rewiring in the topology of E$_2$ treated dataset as well. Indeed, in Fig. 2A and 2B we illustrate how modules alter in size and as seen the modules are highly dynamic in both conditions. In Fig. 3 we provide an example in the case of E$_2$ plus tamoxifen with HDAC1 as seed protein and display the progressive snapshots of the module topology. It has already been indicated that HDAC1 affects breast cancer progression in promoting cell proliferation in association with a reduction in both ER-α protein expression and transcriptional activity [11]. The mid interval module is enriched in Gene Ontology (GO) terms like regulation of nitrogen compound, chromosome organization and regulation of cell proliferation (P-value < 3E-7). Further, we searched for differences in size between the two conditions in the complete time interval (Fig. 2C). Indeed, 33% of our module pairs presented at least two-fold change in their size between the two conditions. Zooming into this subset, we found that in 45% of them, which displayed a three-fold change, the larger module belonged to E$_2$ treatment, while in 7% with fourteen-fold change the larger module belonged to E$_2$ plus tamoxifen treatment. A hypothesis that warrants further study is that the combination of E$_2$ and tamoxifen probably affects

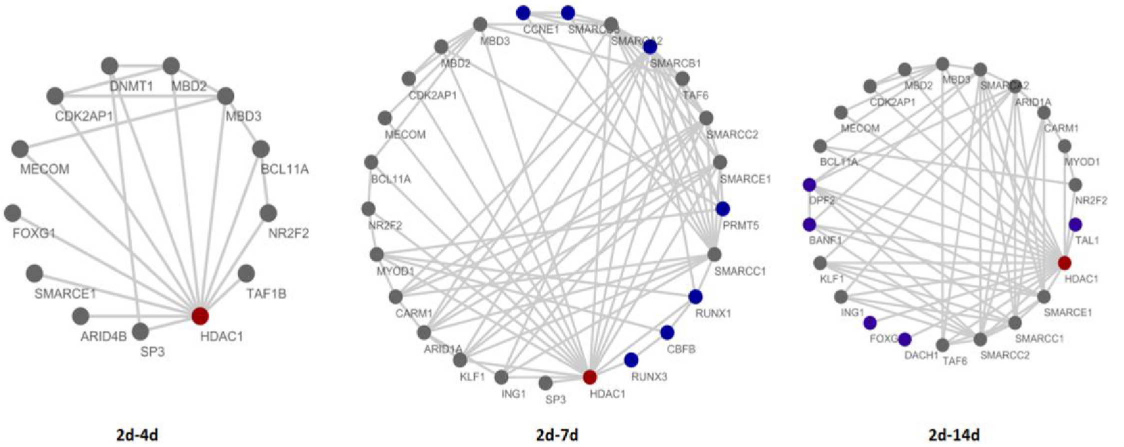

Figure 3.Example of modular rewiring in 3 time intervals (day_2–day_4, day_2–day_7, day_2–day_14) in the E_2 plus tamoxifen treatment. The seed protein (HDAC1) is marked in red and every node that is not member of the previous graph is marked in blue.

the expression behavior of many breast cancer related proteins relative to their interacting partners, inhibiting so the module identification in the topology in comparison to E_2 condition; however, it succeeds in few cases in triggering large paths. A characteristic example is the MCM5 pair of modules. Module construction was detected only in E_2 with members: MCM2, ORC3L, ORC5L, CDC7, CDC6, ORC4L and CDC45 genes. MCM5 and MCM2 are components of the replication fork, which may be responsible for a primary response soon after treatment by reducing DNA regulation [12]. There is also evidence that ORC3L is down-regulated in untreated or permanently tamoxifen treated tumors, CDC7 is over-expressed in multiple cancers, CDC6 is an estrogen responsive gene and CDC45 is up-regulated in proliferating cell populations. Despite the fact that the expression windows between the two conditions are similar, [-2.5, +0.5] and [-2, +0.5] respectively, we observed similar expression trajectories in the E_2 case, whereas the scene changes in E_2 plus tamoxifen treatment, where the expression profiles show dissimilar or even inverse trends in time.

Further in our analysis, we calculated the overlap ratio of members in every pair of modules descending from the same seed protein, in the aforementioned time intervals, to identify the time points in which the modules alter significantly in terms of members between the two conditions. Specifically, we used the overlap ratio metric provided in [13], which was primarily designed for calculating the protein complex coverage of modules. This metric ranges between 0 and 1, with the latter value indicating perfect match. In Fig. 4, we provide the scatter plots that display that day 7 is the one with the highest modular transition.

Moving forward, we compared the graphs (in the complete time interval) that descend after combining the modules of each condition to search for proteins that are absent between them. The E_2 graph included 482 nodes and 1482 edges, whereas the E_2 plus tamoxifen graph 204 nodes and 966 edges. In particular, 329 nodes out of 482 are not included in E_2 plus tamoxifen graph and 45 nodes out of 204 are not included in E_2. We hypothesize that the disruption of these proteins from the topology is translated as follows: either the expression profiles of these proteins changed significantly comparing to their interacting partners, thus they failed to enter the module, or these proteins, due to a broadened functionality repertoire, became members of other modules that were not part of our data. The subset of 329 proteins is enriched in GO terms like cell cycle, regulation of catalytic activity, regulation of cell proliferation, regulation of cell death and cell differentiation, whereas the subset of 45 with terms like system development, cell motion and cell proliferation.

Finally, in Table 1 we present the KEGG pathway analysis of three module pairs with high fold change in size between the two conditions. As seen, the modules are significantly enriched in immune related pathways. Studies have already elucidated the bilateral role of immune system [14]. Cancer cells secrete and respond to cytokines, chemokines and DAMPs influencing the nature and quantity of the immune infiltrate. In order to achieve therapeutic success, any treatment strategy should shift the balance of pro-tumorigenic and anti-tumor immunity in favor of the latter. Interesting observations can be extracted from the LYN and LCK pair of modules, where the E_2 plus tamoxifen module presents an enhanced recruitment of immune response pathways. On the other side, in the case of ERBB2 module pair we found a coupling of ERbB signaling pathway along with T cell receptor, Fc gamma R-mediated phagocytosis and Jak-STAT pathways in E_2 case. The role of ERbB cascade in breast cancer etiology and drug response has long been implicated. Also, there is evidence regarding the role of Jak-STAT cascade in neoplastic transformation and tumor growth. Our findings pose questions about the way the coupling of these pathways is associated with the reported tumor volume increase [9]. Our results, along with other studies that support the immunomodulatory role of tamoxifen [15], corroborate towards the idea that this drug may serve its therapeutic role by affecting major immune signaling pathways.

IV. CONCLUSION

Our integrative approach is a step towards elucidating the dynamic modularity of cellular networks in complex diseases like breast cancer. Our methodology identified a subset of modules that can serve as potential temporal biomarkers of response mechanism to drug activity. Our findings offer a

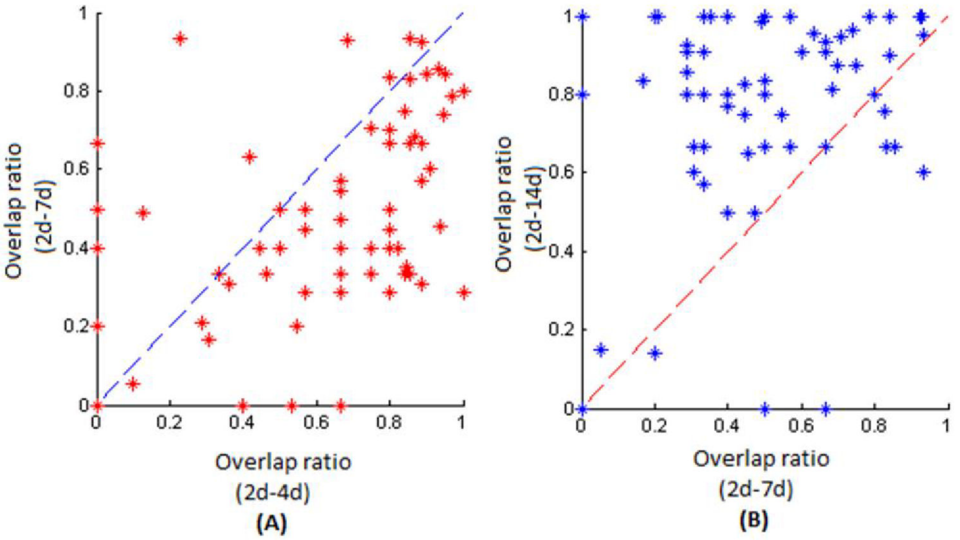

Figure 4. Scatter plots of the overlap ratio, between E_2 and E_2 plus tamoxifen modules, in 3 different time intervals (day_2—day_4, day_2—day_7, day_2—day_14). The dashed line corresponds to the line y = x. As it is evident day 7 is the time point of significant overlap change.

first glimpse of the 'tuning' of protein interplay in time and novel hypotheses for the role of genes/proteins with altered position in the modular organization of interactome topology.

References

[1] J. Li, A. E. G. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima, M. D. O'Connor-McCourt, and E. Wang, "Identification of high-quality cancer prognostic markers and metastasis network modules", *Nat. Commun.*, vol. 1, pp. 1–8, July 2010.

[2] I. W. Taylor, *et al*., "Dynamic modularity in protein interaction networks predicts breast cancer outcome", *Nature Biotechnology*, vol. 27, pp. 199-204, February 2009.

[3] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos, "An in silico method for detecting overlapping functional modules from composite biological networks", *BMC Systems Biology*, 2:93, November 2008.

[4] I. Ulitsky, and R. Shamir, "Identifying functional modules using expression profiles and confidence-scored protein interactions", *Bioinformatics*, vol. 25, pp. 1158-1164, March 2009.

[5] L. Tian, *et al*., "Discovering statistically significant pathways in expression profiling studies", *PNAS*, vol. 102, pp. 13544-13549, September 2005.

[6] H.Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis", *Molecular Systems Biology*, 3:140, October 2007.

[7] E. Georgii, S. Dietmann, T. Uno, P. Pagel, and K.Tsuda, "Enumeration of condition-dependent dense modules in protein interaction networks", *Bioinformatics*, vol. 25, pp. 933-940, February 2009.

[8] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos, "Growing functional modules from a seed protein via integration of protein interaction and gene expression data", *BMC Bioinformatics*, 8:408, October 2007.

[9] K. J. Taylor, A. H. Sims, L. Liang, D. Faratian, M. Muir, G. Walker, B. Kuske, J. M. Dixon, D. A. Cameron, D. J. Harrison, and S. P. Langdon, "Dynamic changes in gene expression in vivo predict prognosis of tamoxifen-treated patients with breast cancer", *Breast Cancer Research*, 12:R39, June 2010.

[10] R. C. Gentleman, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics", *Genome biology*, vol. 5:R80, September 2004.

[11] H. Kawai, *et al*., "Overexpression of histone deacetylase HDAC1 modulates breast cancer progression by negative regulation of estrogen receptor alpha", *Int. J. Cancer*, vol. 107, pp. 353-8.

[12] K. Labib, J. A. Tercero, and J. F. Diffley, "Uninterrupted MCM2-7 function required for DNA replication fork progression", *Science*, vol. 288, pp.1643-1647, June 2000.

[13] X. Wang, Z. Wang, and J. Ye, "HKC: An Algorithm to Predict Protein Complexes in Protein-Protein Interaction Networks", *Journal of Biomedicine and Biotechnology*, vol. 2011, Article ID 480294, August 2011.

[14] R. D. Schreiber, L. J. Old, and M. J. Smyth, "Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion", *Science*, vol. 331, pp. 1565-1570.

[15] S. Behjati, and M. H. Frank, "The effects of tamoxifen on immunity", *Curr. Med. Chem*., vol. 16, pp. 3076-80, July 2009.