Global gene map for cancer reveals pathway hotspots

Pankaj Chopra
Dept. of Computer Science
North Carolina State University
Raleigh, NC 27606
Email: pchopra@ncsu.edu

Han Jun Shin
Dept. of Computer Science
and Engineering
Korea University
Seoul, Korea
Email: magicshan@korea.ac.kr

Jaewoo Kang
Dept. of Computer Science
and Engineering
Dept. of Biostatistics
Korea University
Seoul, Korea
Email: kang.jaewoo@gmail.com

***Abstract*—There have been a number of papers on meta-analysis of microarray datasets. Most of these papers have focused on genes (i.e., differential expression of genes or co-expression of gene pairs) to arrive at a prognostic signature. However, there has been no concerted meta-analysis of expression data in relation to ontologies and pathways. The contribution of this paper is two-fold. First, it uses expression data to create global cancer maps for GO, KEGG and PFAM. These maps reveal hotspots of activation/de-activation. This would be the largest meta-analysis of microarray data in terms of number of datasets and types of cancers represented. Second, in order to prove the concept, we perform an in-depth analysis of the biological processes, pathways and proteins associated with breast cancer. This analysis reveals evidence of a strong link between the GO/KEGG/PFAM hotspots and breast cancer.**

I. Introduction

Over the last few years, there has been an explosion in the number of cancer microarray datasets available in public repositories. However, the number of research papers using multiple datasets in their analysis have been limited [1], [2]. While some papers have attempted to match significantly expressed genes to the Gene Ontology and KEGG pathways, there haven't been many papers that have mapped multiple cancer datasets to GO or KEGG. Mapping and mining multiple microarray datasets may yield insights that were not possible by using just one, or at best, a few datasets.

The results presented in this paper should be viewed with the knowledge that the fold changes linked to 'over-expression' and 'under-expression' are open to biological interpretation. The meta-analysis also suffers from the drawback that the datasets have been taken from diverse platforms and they contain diverse number of genes, and that the datasets themselves have been normalized separately. Nevertheless, we believe that the global cancer map presented here presents some salient pathways that are unique to specific cancers, and also some pathways that are common across cancer types. We have created global cancer maps for Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Protein Families database (PFAM), although here we have just presented the global map for GO.

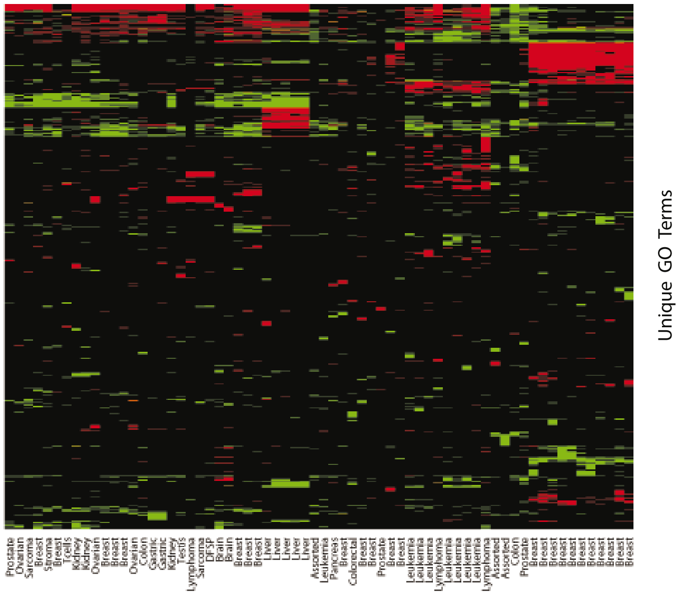

Fig. 1. Global cancer map of significant Gene Ontology (GO) Terms. Each row represents a unique GO term and each column represents a microarray dataset. For a dataset, a red cell represents a GO term found to be significant in relation to the set of over-expressed genes. Similarly, a green cell represents a GO term found to be significant in relation to the set of under-expressed genes

II. Methodology

A total of 67 datasets containing 4,063 cancer tissue samples were downloaded from Gene Expression Omnibus (GEO) and Stanford Microarray Database (SMD). These datasets cover a wide variety of cancer tumors across many microarray platforms. Similar to [1] we scaled the datasets after determining the log2 values of affymetrix datasets (after setting to 16,000 any value that was more than 16,000 and setting to 10 any value that was less than 10). For spotted cDNA datasets, we used the log2 ratio between the measured sample and the control sample. Any gene that had more than 25% values missing was discarded from further analysis. A gene was considered to be over-expressed in a dataset if it showed greater than two-fold increase (i.e., a log2 value greater than 1) in expression levels in more than 80% samples in that dataset.

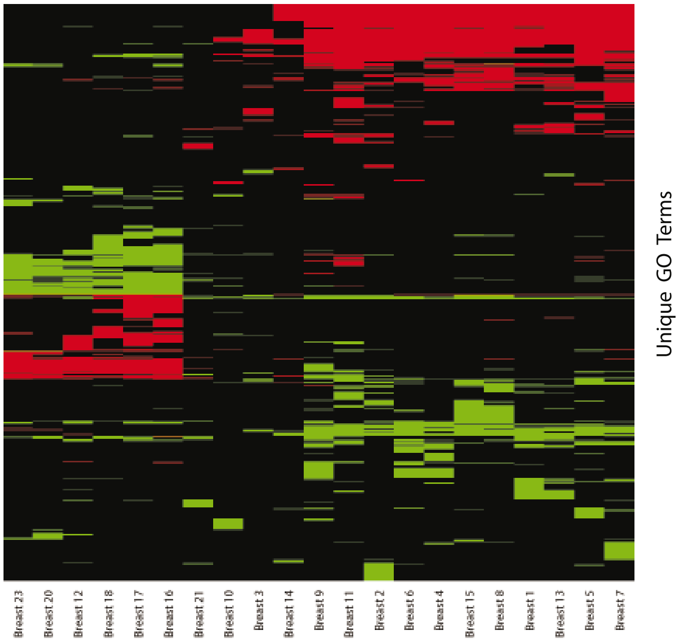

Fig. 2. Breast cancer map of significant Gene Ontology (GO) Terms. Each row represents a unique GO term and each column represents a microarray dataset. For a dataset, a red cell represents a GO term found to be significant in relation to the set of over-expressed genes. Similarly, a green cell represents a GO term found to be significant in relation to the set of under-expressed genes

Similarly, a gene was considered to be under-expressed in a dataset if it showed greater than two-fold decrease (i.e., a log2 value less than -1) in expression levels in more than 80% tumor samples in the dataset. We next calculate the hypergeometric probability of this set of over (under) expressed genes being associated with a particular GO/KEGG/PFAM term.

For the set of over (or under) expressed genes in a dataset, we evaluate if there are any GO/KEGG/PFAM terms that are over-represented, than would be expected by chance. We evaluate this probability by using the hypergeometric distribution of the genes. The probability of a gene set of size *S* containing x genes belonging to a particular GO/KEGG/PFAM term, given that the reference dataset of *N* genes has a total of *A* genes belonging to that particular GO/KEGG/PFAM term is:

$$Pr\{X = x|N, A, S\} = \frac{\binom{A}{x}\binom{N-A}{S-x}}{\binom{N}{S}}$$

where X is a random variable representing the number of over (or under) expressed genes, that are associated with a particular GO/KEGG/PFAM term. A GO/KEGG/PFAM term is considered *significant* only if it has a p-value less than 0.01.

III. Global cancer maps for GO,KEGG and PFAM

The global cancer map showing over and under-expressed GO terms, derived from 67 microarray datasets and 14 cancer types, is shown in Figure 1. The figure shows distinct hotspots of activation/de-activation for several cancer types. Each cancer type appears to have some set of GO terms that are affected by the over (or under) expressed genes. These can be seen most distinctly for liver, leukemia and breast cancer.

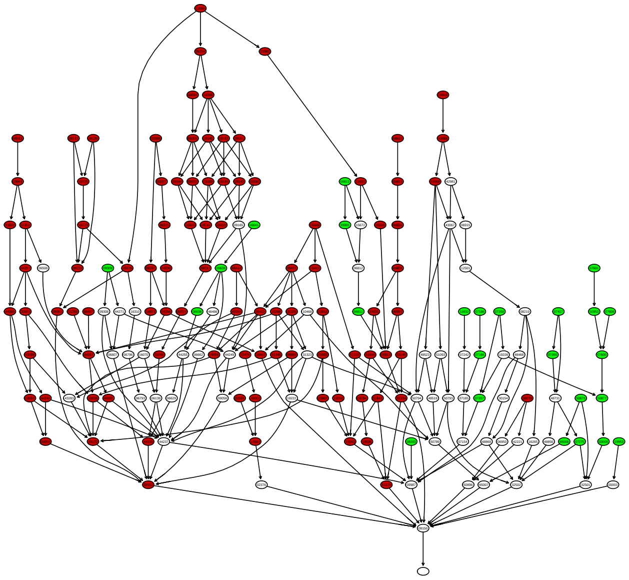

Fig. 3. Cancer map (on GO) for breast cancer datasets. GO Terms associated with over-expressed genes are shown as red nodes, and those associated with under-expressed genes are shown as green nodes.

TABLE I
Top significant KEGG terms for breast cancer datasets

S.No.	KEGG ID	KEGG Term
1	00592	alpha-Linolenic acid metabolism
2	03050	Proteasome
3	04950	Maturity onset diabetes of the young
4	01040	Polyunsaturated fatty acid biosynthesis
5	00062	Fatty acid elongation in mitochondria

TABLE II
Top significant PFAM terms for breast cancer datasets

S.No.	PFAM ID	PFAM Term
1	PF07654	Immunoglobulin C1-set domain
2	PF06758	Repeat of unknown function (DUF1220)
3	PF00572	Ribosomal protein L13
4	PF00240	Ubiquitin family
5	PF00244	14-3-3 protein

We next focus our analysis on significant GO/KEGG/PFAM terms that have been revealed by using only the breast cancer datasets, and show that there is strong evidence to link these specific GO/KEGG/PFAM terms to breast cancer.

IV. Discussion

We created a breast cancer GO map using significant GO terms obtained from breast cancer datasets (Figure 2). We then mapped the significant GO terms associated with over (and under) expressed genes onto the GO tree, as shown in Figure 3. It can be seen that the over (and under) expressed genes in breast cancer target specific GO terms that are distinct from each other. Similarly, we created a breast cancer KEGG map using significant KEGG pathways obtained from breast cancer datasets. We listed the KEGG terms associated with over (and under) expressed genes (p value less than 0.01). Some of the most significant KEGG pathways obtained are shown in Table I. To determine the biological significance of these KEGG terms (Table I), we searched existing literature for links between the KEGG term and breast cancer. Results indicate that there is strong evidence for links between the KEGG terms considered significant from the meta-analysis (Table I) and breast cancer. Links between breast cancer and *alpha-Linolenic acid metabolism* have been reported by [3]. Similarly, associations between breast cancer and *Proteasome* [4]–[6], *Maturity onset diabetes* [7], [8], *Polyunsaturated fatty acid biosynthesis* [9] and *Fatty acid elongation in mitochondria* [10] have been reported.

The most significant PFAM terms associated with over (and under) expressed genes in breast cancer datasets are shown in Table II. Existing evidence [11]–[17] validates a strong connection between the protein families found by our analysis and cancer.

V. Conclusion

We have used a very large number of cancer datasets of various platforms to create global cancer maps for GO, KEGG and PFAM terms. We have focused on breast cancer and have validated that the significant GO/KEGG/PFAM terms from our analysis have biological significance and are strongly linked to cancer. Our future work will focus on three areas. First, to conduct similar analysis for leukemia and liver cancer, for which adequate numbers of microarray datasets are available. Second, to analyze the similarites (and differences) in pathways between these cancers. Third, to incorporate phenotype information into the analysis. This would focus on metastasis (or survival) across cancer types to determine if there are specific pathways associated with particular clinical phenotypes.

References

[1] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer." *Nat Genet*, vol. 36, no. 10, pp. 1090–8, Oct 2004. [Online]. Available: 10.1038/ng1434

[2] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 25, pp. 9309–9314, 2004. [Online]. Available: http://www.pnas.org/content/101/25/9309.full

[3] V. Klein, V. Chajes, E. Germain, G. Schulgen, and M. Pinault, "Low alpha-linolenic acid content of adipose breast tissue is associated with an increased risk of breast cancer." *European Jour of cancer*, vol. 36, no. 3, 2000.

[4] R. Z. Orlowski and C. E. Dees, "The role of the ubiquitination-proteasome pathway in breast cancer: Applying drugs that affect the ubiquitin-proteasome pathway to the therapy of breast cancer," *Breast Cancer Res*, vol. 5, no. 1, 2003.

[5] C. Marx, C. Yau, S. Banwait, Y. Zhou, G. Scott, B. Hann, and J. Park, "Proteasome-regulated erbb2 and estrogen receptor pathways in breast cancer," *Mol. Pharmacol.*, vol. 71, no. 6, 2007.

[6] T. J. Haiming Xu, Donghong Ju and Y. Xie, "Diminished feedback regulation of proteasome expression and resistance to proteasome inhibitors in breast cancer cells," *Breast Cancer Research and Treatment*, vol. 107, no. 2, 2008.

[7] K. B. Michels, C. G. Solomon, F. B. Hu, B. A. Rosner, S. E. Hankinson, G. A. Colditz, and J. E. Manson, "Type 2 Diabetes and Subsequent Incidence of Breast Cancer in the Nurses' Health Study," *Diabetes Care*, vol. 26, no. 6, pp. 1752–1758, 2003.

[8] A. H. Wu, M. C. Yu, C.-C. Tseng, F. Z. Stanczyk, and M. C. Pike, "Diabetes and risk of breast cancer in Asian-American women," *Carcinogenesis*, vol. 28, no. 7, pp. 1561–1566, 2007.

[9] J. Menendez and R. Colomer, "Inhibition of fatty acid synthase-dependent neoplastic lipogenesis as the mechanism of ?-linolenic acid-induced toxicity to tumor cells: an extension to nwankwos hypothesis ." *Medical Hypothese*, vol. 64, no. 2, 2005.

[10] F. Kuhajda, "Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology," *Nutrition*, vol. 16, no. 3, 2000.

[11] M. Cicardi, A. Beretta, M. Colombo, D. Gioffre, M. Cugno, and A. Agostoni, "Relevance of lymphoproliferative disorders and of anti-c1 inhibitor autoantibodies in acquired angio-oedema," *Clinical & Experimental Immunology*, vol. 106, no. 3, 2003.

[12] D. B. Rubinstein, A. Stortchevoi, M. Boosalis, R. Ashfaq, and T. Guillaume, "Overexpression of DNA-binding Protein B Gene Product in Breast Cancer as Detected by in Vitro-generated Combinatorial Human Immunoglobulin Libraries," *Cancer Res*, vol. 62, no. 17, pp. 4985–4991, 2002. [Online]. Available: http://cancerres.aacrjournals.org/cgi/content/abstract/62/17/4985

[13] Y. Sasaki, T. Kobayashi, H. Mita, M. Toyota, H. Suzuki, K. Imai, Y. Shinomura, and T. Tokino, "Activation of the ribosomal protein L13 gene in human gastrointestinal cancer," *AACR Meeting Abstracts*, vol. 2006, no. 2, pp. A22–, 2006.

[14] A. Amsterdam, K. Sadler, K. Lai, S. Farrington, and R. Bronson, "Many ribosomal protein genes are cancer genes in zebrafish," *PLoS Biology*, vol. 2, no. 5, 2004.

[15] T. Ohta and M. Fakuda, "Ubiquitin and breast cancer," *Oncogene*, vol. 23, no. 11, 2004.

[16] R. Kobayashi, M. Deavers, R. Patenia, T. Rice-Stitt, J. Halbe, S. Gallardo, and R. Freedman, "14-3-3 zeta protein secreted by tumor associated monocytes/macrophages from ascites of epithelial ovarian cancer patients." *Cancer Immunol Immunother.*, 2008.

[17] J. M. A. Moreira, G. Ohlsson, F. E. Rank, and J. E. Celis, "Down-regulation of the Tumor Suppressor Protein 14-3-3sigma Is a Sporadic Event in Cancer of the Breast," *Mol Cell Proteomics*, vol. 4, no. 4, pp. 555–569, 2005.