Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 09-029

Integrative Biomarker Discovery for Breast Cancer Metastasis from Gene Expression and Protein Interaction Data Using Error-tolerant Pattern Mining

Rohit Gupta, Smita Agrawal, Navneet Rao, Ze Tian, Rui Kuang, and Vipin Kumar

November 24, 2009

Integrative Biomarker Discovery for Breast Cancer Metastasis from Gene Expression and Protein Interaction Data Using Error-tolerant Pattern Mining

Rohit Gupta[1] rohit@cs.umn.edu
Smita Agrawal[2] agraw034@umn.edu
Navneet Rao[1] nrao@cs.umn.edu
Ze Tian[1] tianze@cs.umn.edu
Rui Kuang[1] kuang@cs.umn.edu
Vipin Kumar[1] kumar@cs.umn.edu

[1]Computer Science and Engineering, University of Minnesota, Minneapolis MN 55455 USA
[2]Lab Medicine and Pathology, University of Minnesota, Minneapolis MN 55455 USA

Abstract

Biomarker discovery for complex diseases is a challenging problem. Most of the existing approaches identify individual genes as disease markers, thereby missing the interactions among genes. Moreover, often only single biological data source is used to discover biomarkers. These factors account for the discovery of inconsistent biomarkers. In this paper, we propose a novel error-tolerant pattern mining approach for integrated analysis of gene expression and protein interaction data. This integrated approach incorporates constraints from protein interaction network and efficiently discovers all patterns (groups of genes) in a bottom-up fashion from the gene-expression data. We call these patterns active sub-network biomarkers. To illustrate the efficacy of our proposed approach, we used four breast cancer gene expression data sets and a human protein interaction network and showed that active sub-network biomarkers are more biologically plausible and genes discovered are more reproducible across studies. Finally, through pathway analysis, we also showed a substantial enrichment for known cancer genes and hence were able to generate relevant hypotheses for understanding the molecular mechanisms of breast cancer metastasis.

1. INTRODUCTION

Most of the complex problems like biomarker discovery require more information than provided by any individual biological data. For example, since both gene expression data and protein interaction data are noisy, biomarkers identified using information from both gene expression and protein interaction data are more reliable and biologically plausible than those obtained from individual data sources. This is because a set of genes that co-express as well as physically interact with each other is more likely to be significant and biologically relevant. However, most of the previous studies rely on single biological data, for example, gene expression or protein interaction data to find disease biomarkers. Another issue is that most approaches identify individual genes as disease markers through analysis of genome-wide expression profiles [1, 2] and therefore ignore the interactions among them.

There are several problems with both, individual gene-based biomarkers and group of genes as biomarkers obtained from a single biological data source: 1) Poor reproducibility – no or very little overlap among biomarkers across various studies. For example, for the problem of breast cancer metastasis, Van de veer et al [3] identified 70 gene signatures and Wang et al [4] identified 76 gene signatures, however surprisingly, only 3 overlap; 2) Poor interpretability – difficult to understand the underlying mechanism of a gene signature when genes have either unknown functions or may be associated with unrelated biological pathways; 3) Poor coverage – some of the relevant genes can be missed in differential analysis. For example, in the breast cancer case, known cancer genes, such as P53, KRAS, etc, may not be detected [5]; 4) Poor predictive capability – biomarkers identified using single biological information tend to have inferior predictive power [5, 6].

To address some of the limitations mentioned above, a number of studies have been proposed that use multiple sources of information [7-20]. There are several pathway based approaches [21-24] that score pathways by observing the coherency of the gene expression values among the pathway genes. However, these approaches are limited because a majority of genes have not yet been assigned to a pathway. Moreover, not all genes in a pathway may be responsive for the phenotype under study [6]. Another set of approaches use a complete protein interaction network and try to extract relevant sub-networks based on coherent expression patterns of their genes. One of the promising works [5] in this category uses a greedy algorithm and starts from a random gene to find sub-networks whose constituent genes co-express with each other. Although this is a promising approach, since it employs heuristic-based greedy algorithm, it cannot guarantee completeness. Another interesting study [25] integrates gene expression and protein interaction data to enumerate dense modules with the provision of integrating additional constraints from a variety of data sets. This approach is primarily designed for finding protein complexes from protein interaction data. Hence, first of all, it is not directly applicable to case-control type of data and secondly, it is sensitive to noise in the gene expression data and protein interaction network as it only looks for dense modules.

To address the above issues, building on our recent work [26], we propose a systematic error-tolerant pattern mining based approach, which sequentially generates all patterns that satisfy the user constraints in a bottom-up

fashion. The application of error-tolerant pattern mining for discovering biomarkers from multiple biological data sources is a novel framework and to the best of our knowledge, has not been explored before. Unlike traditional association mining algorithms, this approach directly works on real-valued attributes and does not require binarization of the gene expression data, which results in loss of information. Moreover, it directly accounts for errors/noise in the data and discovers patterns with user-defined error tolerance.

In addition, we used an anti-monotonic measure 'diameter' defined on protein interaction network to find patterns (groups of genes) that not only show coherent expression values but whose constituent set of genes are connected in the protein interaction network. We call these resultant patterns active sub-network biomarkers as the constituent genes are discriminative of the two groups of samples; are over- or under-expressed for at least some of the samples; and induce a sufficiently connected sub-graph in the protein interaction network. For this reason, we will use patterns and active sub-network biomarkers interchangeably in the rest of the paper. In contrast with the other approaches, our proposed method addresses both the issues of completeness and errors/noise in the gene-expression data.

To illustrate the efficacy of our proposed pattern mining based approach, we used four case-control breast cancer gene-expression data sets and a human protein interaction network. An integrated analysis is performed to discover all active sub-network biomarkers for breast cancer metastasis. We showed that the genes discovered as part of the biomarkers are not only more reproducible across data sets, these biomarkers themselves are enriched for many known gene sets obtained from MSigDB [23]. We also performed pathway analysis (using IPA software) on the identified genes and showed a substantial enrichment for known cancer genes. In addition, we were able to generate relevant hypotheses for understanding the molecular mechanisms of breast cancer metastasis. Overall, the results presented in this study strongly suggest that our proposed error-tolerant pattern mining approach is a potential integrative method to discover active sub-network biomarkers.

1.1 Contributions

Following are the key contributions of the paper:

- Systematic approach to exhaustively enumerate all patterns or active sub-network biomarkers that satisfy the user-specified criteria.
- Demonstration of the reproducibility of genes discovered as part of biomarkers across four breast cancer data sets.
- Comprehensive biological evaluation of the discovered patterns using MSigDB-based enrichment analysis and of the discovered genes using pathway/network analysis (IPA software). Also, some relevant hypotheses for understanding the molecular mechanisms underlying breast cancer metastasis were generated.

2. ACTIVE SUB-NETWORK BIOMARKER DISCOVERY

We recently proposed a pattern mining approach to efficiently discover all constant rows/columns error-tolerant patterns from a real-valued gene expression data [26]. The input to this algorithm includes the gene expression data and various user-defined parameters such as *range* (measures the coherence of expression values), *SampleSupport* (minimum number of samples supporting the pattern), *RangeSupport* (measures the contribution of samples toward a particular pattern), row and column error-tolerances (maximum number of errors allowed in the pattern). The output of this algorithm is the set of all patterns (groups of genes and their supporting samples) that satisfy the above user-defined constraints. The genes in each pattern are expected to co-express for the samples supporting the pattern. We show in figure 1, three examples of error-tolerant patterns, namely A, B and C. Please note that pattern A has one error in each row 1 and 3, while patterns B and C have no error values. The efficient discovery of all patterns from the expression data was made possible due to systematic pruning of search space. This was done using a noise-tolerant support function and the anti-monotonic constraints (refer to [26] for details). This approach has several advantages over biclustering approaches (refer to [27, 28] for survey) and other previously published error-tolerant pattern mining approaches (refer to [29] for survey). First, this approach directly works on the real-valued attributes and hence eliminates the need to binarize the data, which often results in the loss of information. Second, it tolerates errors in the patterns, which is crucial in order to discover true underlying patterns that are fragmented due to noise/errors in the data. Third, it uses anti-monotonic constraints and thereby allows the efficient discovery of all the patterns. Finally, it provides a natural framework for integrating other data sources as anti-monotonic constraints and therefore as a result produces patterns that are inferred from multiple data sources.

In this paper, we extended the above error-tolerant pattern mining approach for an integrated analysis of gene expression data and protein interaction network. One of the strategies to perform such an analysis is to derive an anti-monotonic measure on protein interaction network and incorporate it in the pattern mining process on the gene-expression data. For this, we defined a measure called 'diameter', which for any sub-network is defined as

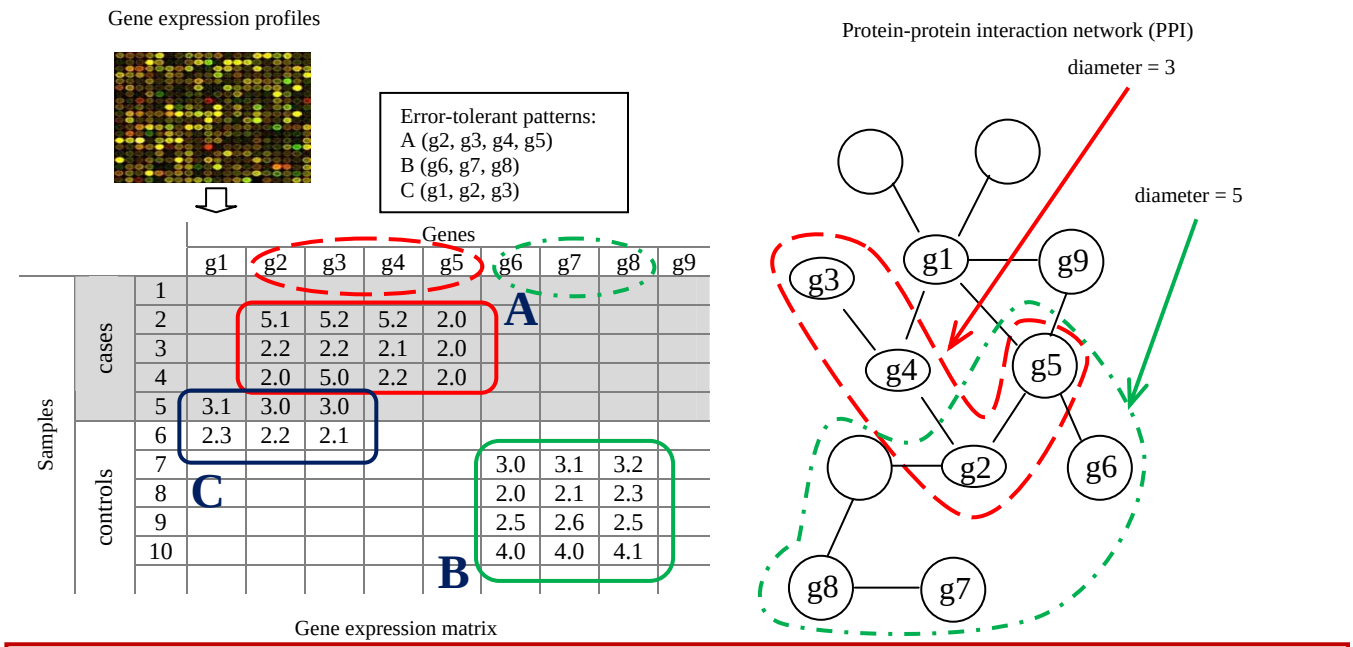

Figure 1: Schematic overview and an illustrative example of active sub-network biomarker

the maximum of the shortest pair-wise distances of all the constituent proteins computed using the complete graph. As can be observed, this measure is anti-monotonic as it can only increase on the addition of any protein to the sub-network. For integrated analysis, we imposed this constraint in addition to the ones used for the analysis of gene-expression data alone and obtained patterns, i.e. groups of genes that not only co-express but their proteins interact as well. Consider for example figure 1. Although patterns A, B and C are all valid error-tolerant patterns, pattern A has a diameter of 3 and pattern B has a diameter of 5 and hence for a diameter threshold of 3, pattern B will be pruned.

Finally, to discover biomarkers, using the case-control class label, we only select those patterns that have a p-value less than 0.05 and odds ratio of more than 2.0 (pattern more represented in cases) or less than 0.5 (pattern more represented in controls). For example in figure 1, pattern C will be discarded in this step even though it may be a valid error-tolerant pattern. This is because it occurs equally in both cases and controls and hence is not discriminative.

3. RESULTS AND DISCUSSION

In this section, we present our experimental methodology and results to demonstrate the efficacy of our proposed approach in discovering statistically significant and biologically meaningful disease biomarkers from an integrated analysis of gene-expression and protein interaction data.

3.1 Data sets and pre-processing

We used four breast cancer gene-expression data sets, all of which were taken from Affymetrix platform HG-U133A. In addition, we used human protein interaction network [5], which comprises of 57,235 interactions among 11,203 proteins. All the four breast cancer data sets were downloaded from GEO website: Desmedt (GSE7390), Loi (GSE6532), Miller (GSE3494) and Pawitan (GSE1456). There were a total of 22,283 probe sets and after removing the ones that have no associated gene symbol or more than one associated gene symbols, we grouped the probe sets with the same gene symbol and ended up with 12,733 unique genes. The expression value of each gene is computed by taking the average of the corresponding probe sets. Furthermore, only 8920 of the 12,733 genes that can be mapped to the protein interaction network were considered for final analysis. All the four resulting gene expression data sets were finally normalized using RMA-normalization approach. The patient samples in the four datasets are classified as cases and controls based on their metastasis state. The patients who developed metastasis within 5 years of prognosis were considered as metastasis cases. The patients who were free of metastasis longer than 8 years of survival and follow-up time were considered as controls. The case-control ratio for Desmedt, Loi, Miller and Pawitan data

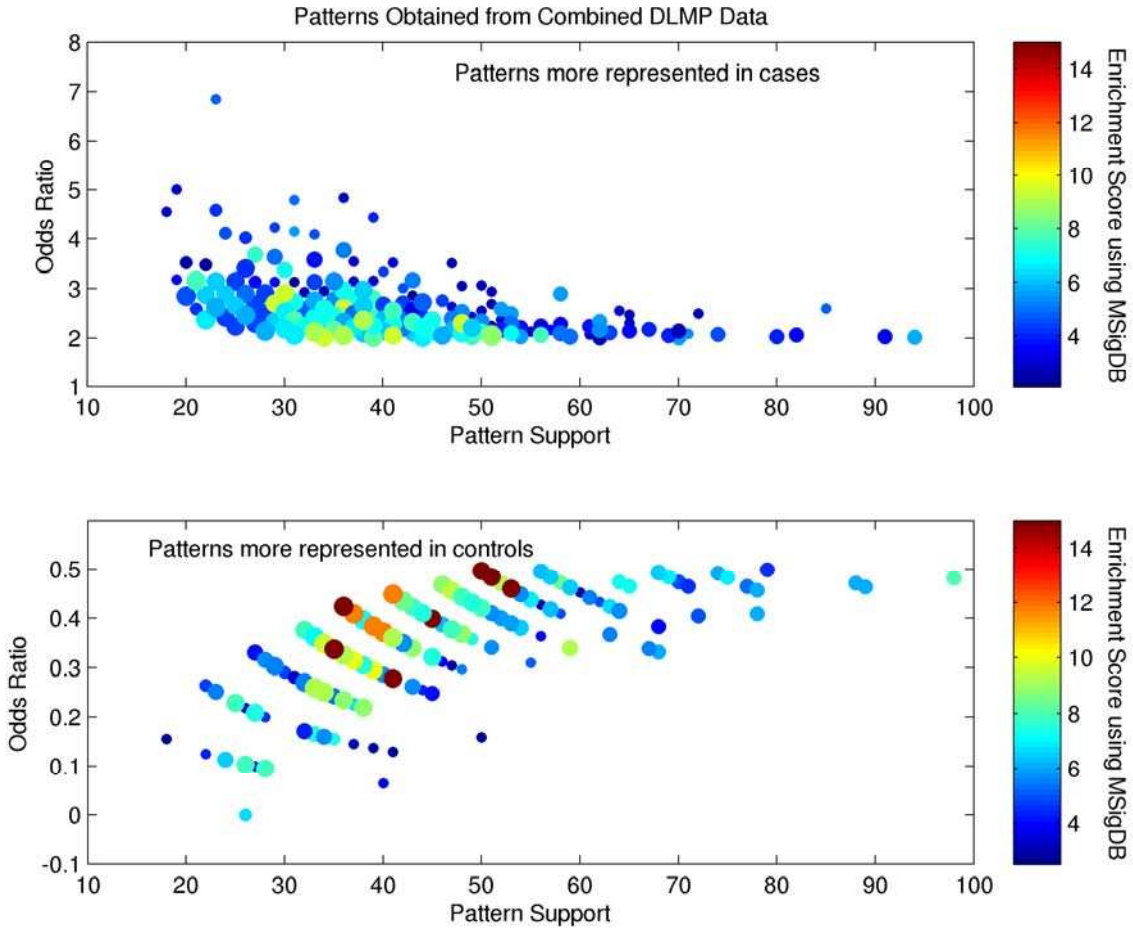

Figure 2: (Best viewed in color) Active sub-network biomarkers obtained from combined DLMP data. (Marker color indicates the enrichment score and marker size indicates the size of the pattern).

set was 35:136, 51:112, 37:150 and 35:35 respectively.

3.2 Active sub-network biomarkers are enriched based on MSigDB gene sets

In the first experiment, to increase the sample size, we combined all the four breast cancer gene-expression data sets and discovered active sub-network biomarkers on the combined data (we refer to it as DLMP data) using our proposed pattern mining based approach. A total of 1,777 patterns (active sub-network biomarkers) were identified, with pattern sizes (number of genes in them) as large as 8. Please note we used a 'diameter' threshold of 3 and error-tolerances (both row and column) of 0.25 for this experiment. Since the discovered patterns essentially represent groups of genes, we evaluated them using the enrichment analysis based on gene sets obtained from Molecular Signature Database (MSigDB) [23]. This entails finding a p-value for each (pattern, MSigDB gene set) pair, which determines the statistical significance of the discovered patterns. Enrichment score for each pattern is then computed as $-\log_{10}(\text{p-value}_{min})$. Similarly, enrichment score for each MSigDB gene set is computed. Figure 2 shows the odds ratio (shown on the y-axis), support (number of samples supporting the pattern – shown on the x-axis), enrichment score based on MSigDB (shown as marker color) and pattern size (shown as marker size) of all the 1,777 patterns. It can be observed that some of the big patterns (6 or 7 constituent genes) have enrichment scores as high as 9 and odds ratio above 2 indicating over- or under-expression of constituent genes in cases. Similarly, some patterns have enrichment scores as high as 15 and odds ratio below 0.5 indicating over- or under-expression of constituent genes in controls. Moreover, to give an overall statistics, 87.67%, 74.17% and 54.36% of the patterns were enriched with at least one MSigDB gene set when p-value thresholds of 1e-3, 1e-4 and 1e-5 were used respectively. For the same p-value thresholds, 12.87%, 7.3% and 3.18% gene sets enriched at least one pattern. It is noteworthy that most of

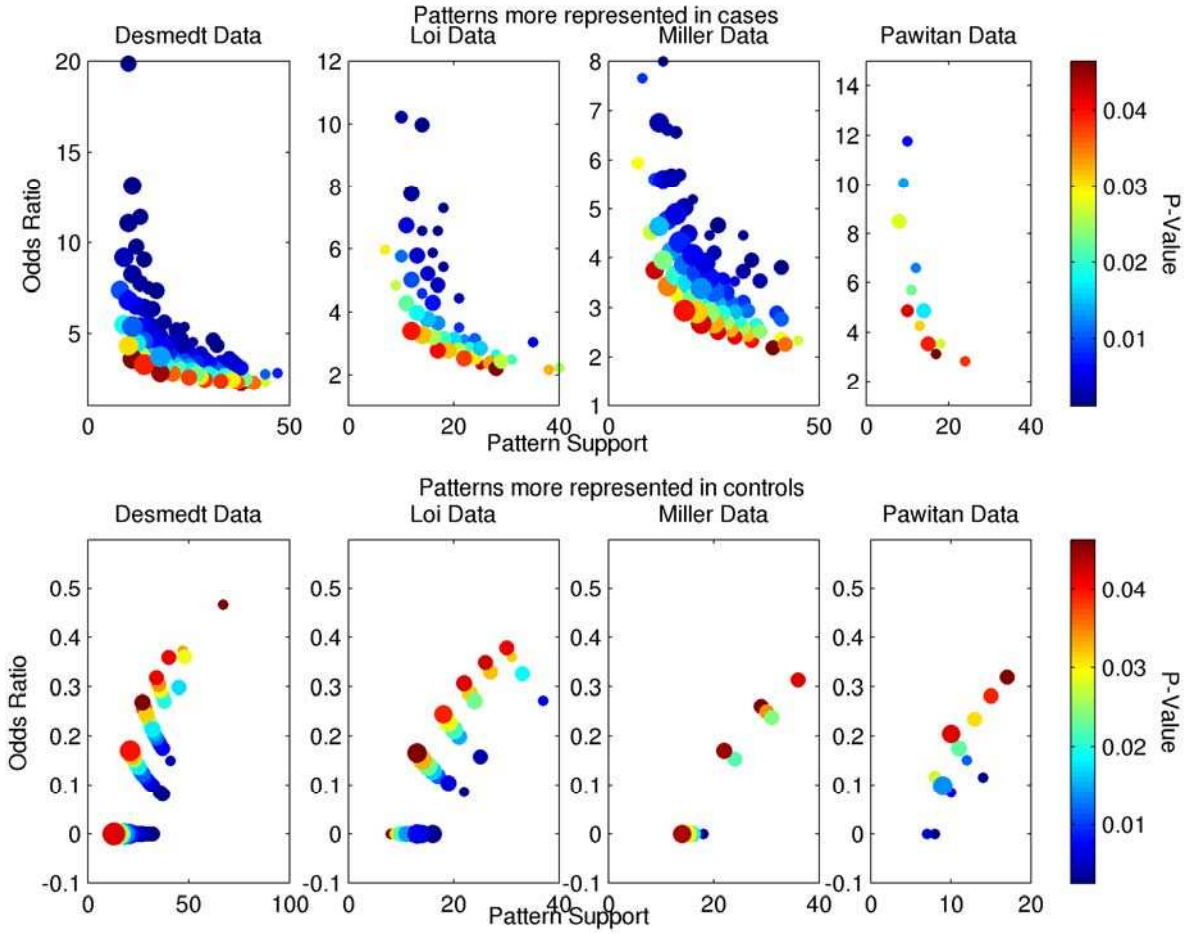

Figure 3: (Best viewed in color) Active sub-network biomarkers obtained from four breast cancer data sets. (Marker color indicates the empirical p-value and marker size indicates the size of the pattern).

these gene sets were manually curated from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts that includes many cancer related gene sets.

3.3 Increased reproducibility of genes identified from active sub-network biomarkers across data sets

Next, we examined the agreement in the genes identified from four different data sets using our proposed pattern mining approach. To do this, using the same 'diameter' and error-tolerance parameters as used in the previous experiment, we applied our approach to discover all patterns or active sub-network biomarkers from each of the four breast cancer data sets. Please note that same protein interaction network was used in each case and only those patterns were considered whose odds ratio were either less than 0.5 (more represented in controls) or greater than 2 (more represented in cases) with a p-value less than 0.05. We identified a total of 6266 (covering 199 genes), 580 (covering 166 genes), 1711 (covering 157 genes) and 111 (covering 82 genes) active sub-network biomarkers from Desmedt, Loi, Miller and Pawitan data respectively. The odds ratio, support, p-value (shown as marker color) and size (shown as marker size) of these patterns is shown in figure 3 for each of the data sets. Overall, 274 unique genes were identified from these 4 independent analyses and interestingly, 52 of these genes were identified from all the four data sets, 56 were identified from three, 62 were identified from two, and finally 104 genes were identified from only one of the four data sets.

On the other hand, if we select individual marker genes from these four data sets separately using the same odds ratio and p-value thresholds and without incorporating protein interaction network information, only 46 unique genes were identified. It is noteworthy that

out of these, only 1 gene was identified from two data sets while others were identified from only one of the four data sets. These results suggest that individual genes have very limited information and finding groups of genes as biomarkers while incorporating protein interaction network information yields more consistent and meaningful results.

One question that may arise is: whether this increased overlap among the genes is due to the network bias? There are several ways to answer this question. First, random sub-networks, equal in number and size of the discovered active sub-network biomarkers, can be extracted from the protein interaction data and then overlap among them can be observed. Another possible way is to use the same protein interaction network as prior biological knowledge but randomize the gene expression data. The application of the proposed integrated approach on this expression data should hypothetically produce biomarkers without significant overlap. Finally, one can also evaluate the biological significance of the discovered genes with respect to the phenotype of interest and it is likely that overlapping genes that also show relevant biological associations with the phenotype are not obtained due to network bias. Due to space limitations, we only show the biological relevance of the discovered genes. Although, using the randomization test explained as the first strategy above, it was shown in [5] that the increased overlap in sub-network based biomarkers is not due to the network bias.

3.4 Biological relevance of discovered genes

Using the IPA (Ingenuity Pathway Analysis) software, we assessed the biological relevance of the 274 unique genes contained in active sub-network biomarkers discovered from the four breast cancer data sets. The top molecular and cellular functions of these genes included cellular movement, cell death, cell growth and proliferation, cell-to-cell signaling interactions and cellular assembly and organization, all of which are processes implicated in the progression of cancer [30]. Cancer stood out as the top biological function/disease enriched, with as many as 172 out of the 274 genes implicated. Out of these, 56 genes were specific to various aspects of breast cancer including proliferation, apoptosis, invasion, migration, survival etc.

The top network obtained as shown in figure 4 is associated with cancer and genetic disorders. Most of the genes shown in this network are contained in our discovered active sub-network biomarkers. In fact, many genes were common in active sub-network biomarkers that were obtained from multiple data sets (shown as dark red colored nodes in the figure). One of the key genes in the network, *ERBB2* (a cell surface receptor), is a known breast cancer gene [31] and its over-expression / amplification has been correlated with an aggressive spread of cancer [32]. Notably, three different collagen genes (*Col5A1*, *Col6A3* and *Col1A1*) and several other extracellular matrix (ECM) signaling genes are associated with *ERBB2* in this network. Previous studies have suggested that the interaction between cell surface proteins on breast cancer cells with ECM components plays an important role in breakdown of the ECM leading to metastasis of cancer cells [33]. Taken together, this suggests that the network components surrounding *ERBB2* discovered as part of our active sub-network biomarkers may be potential target genes involved in breast cancer metastasis and should be further studied.

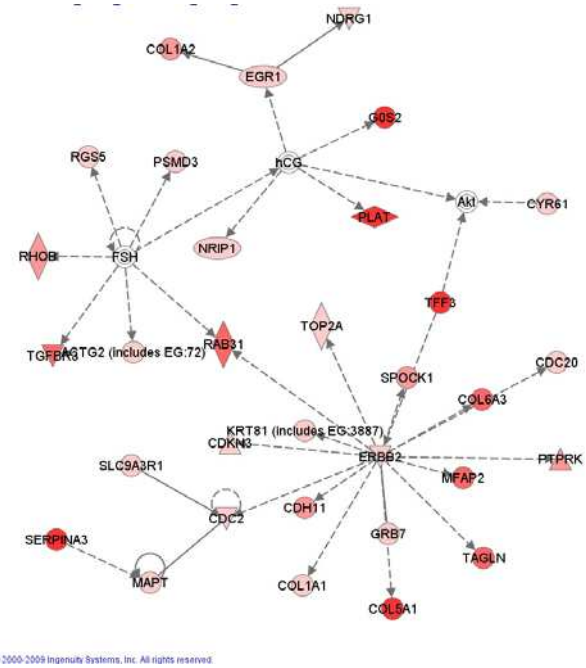

Figure 4: Top biological network obtained through analysis of 274 unique genes. Color shade indicates the number of data sets from which patterns containing that gene were obtained. (dark red genes were obtained from all 4 datasets, whereas light pink genes were obtained from a single dataset).

Interestingly, through this network analysis, we identify *hCG* (human chorionic gonadotropin) and *FSH* (Follicle Stimulating Hormone) as potentially important genes by virtue of the fact that all the genes surrounding them in this network are contained in the identified active sub-network biomarkers. The presence of elevated levels of *hCG* during pregnancy has recently been linked to a protective effect against breast cancer [34]. Thus, studying the role of *FSH*, another fertility hormone, in breast cancer spread and progression may be important.

Additionally, several other networks involved in cancer were identified. Notably, another top network (figure not shown) contains *TP53* (a known cancer gene) and *LGALS3* (a galactose-specific lectin important in ECM interactions). *LGALS3* has been implicated in tumor

progression and its expression has been shown to be up-regulated in breast cancer and metastatic tissue relative to normal breast tissue [35].

4. CONCLUSIONS

We proposed a novel error-tolerant pattern mining approach for the integrated analysis of gene expression data and protein interaction network in order to efficiently discover all active sub-network biomarkers of breast cancer metastasis. Through experiments on four real breast cancer data sets, we showed that this proposed approach has potential as it not only produces consistent biomarkers but those biomarkers are significantly enriched for MSigDB gene sets. Compared to individual gene markers, we showed that our active sub-network biomarkers discovered by integrated analysis are more biologically plausible, more reproducible, and finally more likely to be true than random.

Finally the work presented in this paper can be extended in several ways. In the future, we would like to test the efficacy of active sub-network biomarkers for the classification task. We also plan to apply the proposed approach to several other domains including lung cancer and prostate cancer.

5. REFERENCES

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, and X. Yu, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531, 1999.

[3] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, vol. 347, pp. 1999-2009, 2002.

[4] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671-9, 2005.

[5] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Mol Syst Biol*, vol. 3, pp. 140, 2007.

[6] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Comput Biol*, vol. 4, pp. e1000217, 2008.

[7] X. Chen and L. Wang, "Integrating biological knowledge with gene expression profiles for survival prediction of cancer," *Journal of Computational Biology*, vol. 16, pp. 265-278, 2009.

[8] M. Francesconi, D. Remondini, N. Neretti, J. Sedivy, L. Cooper, E. Verondini, L. Milanesi, and G. Castellani, "Reconstructing networks of pathways via significance analysis of their intersections," *BMC bioinformatics*, vol. 9, pp. S9, 2008.

[9] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Computational Biology*, vol. 4, 2008.

[10] T. H. Hwang, Z. Tian, R. Kuang, and J. P. Kocher, "Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction," presented at Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 2008.

[11] C. Kim, J. Choi, and S. Yoon, "Microarray Data Analysis of Perturbed Pathways in Breast Cancer Tissues," *Genomics Inform*, vol. 6, pp. 210-222, 2008.

[12] R. K. Nibbe, S. Markowitz, L. Myeroff, R. Ewing, and M. R. Chance, "Discovery and Scoring of Protein Interaction Subnetworks Discriminative of Late Stage Human Colon Cancer," *Molecular & Cellular Proteomics*, vol. 8, pp. 827, 2009.

[13] W. Pan, "Network-based model weighting to detect multiple loci influencing complex diseases," *Human Genetics*, vol. 124, pp. 225-234, 2008.

[14] R. B. Russell and P. Aloy, "Targeting and tinkering with interaction networks," *Nature Chemical Biology*, vol. 4, pp. 666-673, 2008.

[15] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nat Biotechnol*, vol. 27, pp. 199-204, 2009.

[16] Z. Tian, T. H. Hwang, and R. Kuang, "A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge," *Bioinformatics*, vol. 25, pp. 2831, 2009.

[17] I. Ulitsky, R. M. Karp, and R. Shamir, "Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles," *Lecture Notes in Computer Science*, vol. 4955, pp. 347, 2008.

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

[19] Z. Yanni, S. Xiaotong, and P. Wei, "Network-based support vector machine for classification of microarray samples," *BMC Bioinformatics*, vol. 10.

[20] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC bioinformatics*, vol. 10, pp. S21, 2009.

[21] P. Pavlidis, D. Lewis, and W. Noble, "Exploring gene expression data with class scores," presented at Pacific Symposium on Biocomputing, Hawaii, 2002.

[22] P. Pavlidis, J. Qin, V. Arango, J. J. Mann, and E. Sibille, "Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex," *Neurochemical Research*, vol. 29, pp. 1213-1222, 2004.

[23] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A*, vol. 102, pp. 15545-50, 2005.

[24] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 13544, 2005.

[25] 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. 933-40, 2009.

[26] Rohit Gupta, Navneet Rao, and Vipin Kumar, "A Novel Error-Tolerant Frequent Itemset Model for Binary and Real-Valued Data," University of Minnesota - Computer Science and Engineering 09-026, 2009.

[27] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Transactions on computational Biology and Bioinformatics*, pp. 24-45, 2004.

[28] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, pp. 1122, 2006.

[29] Rohit Gupta, Gang Fang, Blayne Field, Michael Steinbach, and Vipin Kumar, "Quantitative evaluation of approximate frequent pattern mining algorithms," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas, Nevada, USA: ACM, 2008.

[30] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57-70, 2000.

[31] C. R. King, M. H. Kraus, and S. A. Aaronson, "Amplification of a novel v-erbB-related gene in a human mammary carcinoma," *Science*, vol. 229, pp. 974-6, 1985.

[32] F. Revillion, J. Bonneterre, and J. P. Peyrat, "ERBB2 oncogene in human breast cancer and its clinical significance," *Eur J Cancer*, vol. 34, pp. 791-808, 1998.

[33] Y. A. DeClerck, A. M. Mercurio, M. S. Stack, H. A. Chapman, M. M. Zutter, R. J. Muschel, A. Raz, L. M. Matrisian, B. F. Sloane, A. Noel, M. J. Hendrix, L. Coussens, and M. Padarathsingh, "Proteases, extracellular matrix, and cancer: a workshop of the path B study section," *Am J Pathol*, vol. 164, pp. 1131-9, 2004.

[34] M. K. Johana E. Vanegas, Julia S. Pereira, Hilal Kocdor, Jose Russo, Kara Snider, Fathima Sheriff, Irma H. Russo, "Mammary cancer prevention by short treatment with human chorionic gonadotropin (abstract nr 2059)," presented at 100th Annual Meeting American Association for Cancer Research, 2009.

[35] M. A. Mayoral, C. Mayoral, A. Meneses, L. Villalvazo, A. Guzman, B. Espinosa, J. L. Ochoa, E. Zenteno, and J. Guevara, "Identification of galectin-3 and mucin-type O-glycans in breast cancer and its metastasis to brain," *Cancer Invest*, vol. 26, pp. 615-23, 2008.