An Integrative Approach for Identification of Network Biomarkers in Breast Cancer Subtypes

Forough Firoozbakht, Iman Rezaeian, Alioune Ngom and Luis Rueda

School of Computer Science, University of Windsor
401 Sunset Ave., Windsor, ON, N9B3P4, Canada
{firoozb,rezaeia,angom,lrueda}@uwindsor.ca

Abstract. Breast cancer is a complex disease that has been characterized into ten different molecular subtypes. Current computational methods for determining the subtypes are based on identifying gene biomarkers; i.e. differentially expressed genes that best separate the subtypes. Such methods do not take into account the functional relationships between genes, and hence, may not yield informative biomarkers. We propose a machine learning framework for identifying network biomarkers of breast cancer subtypes; i.e. subnetworks of functionally related to gene biomarkers that best distinguish the subtypes. Our framework incorporates genomics, transcriptomics and interactomics information in identifying discriminative network biomarkers corresponding to each subtype. We applied our method on the METABRIC data and obtained a collection of highly predictive network biomarkers with AUC performances ranging from 89.6% to 99.1%.

Keywords: breast cancer; molecular sub-typing; gene expression; protein interaction network; copy number aberration/variation; network biomarker.

1 Introduction

Breast cancer (BC) is one of the leading causes of cancer related deaths among women worldwide [4]. It has been categorized into different subtypes that can be distinguished based on gene expression characteristics [3]. Correct diagnosis of a patient's specific BC subtype is vital in subsequently determining the best care for the patient. Most bioinformatic methods have focused on identifying BC biomarkers as small subsets of differentially expressed (DE) genes between subtypes. However, DE genes have limited predictive performance due to (i) the tumor heterogeneity within tissues and across patients and (ii) the independence between identified DE genes. New methods aim to alleviate these limitations by integrating additional biological information with gene expression information and identify network biomarkers (NBs); i.e. subnetworks of functionally related to DE genes that best distinguish BC subtypes [7].

2 Materials and Methods

2.1 Data Set

We use the discovery data of METABRIC dataset [3] containing the copy number values and gene expression levels of 997 primary breast tumors with long-term clinical follow-up. Each sample contains expression information of 48,803 probe IDs, which have been mapped to 24,351 unique genes using median expression. The number of samples corresponding to each subtype are listed in Table 1. We combine human protein-protein interaction (PPI) network data from BioGrid [9], HPRD [8], Intact [5] and MINT [2] into a single large PPI network consisting of 230,000 protein-protein interactions and 15,823 proteins; which yield relationship information between genes.

Table 1. Number of samples in each of the ten subtypes.

Subtypes	1	2	3	4	5	6	7	8	9	10
# of Samples	76	45	156	167	94	44	109	143	67	96

2.2 Approach

We propose a new method for identifying the NB specific to each BC subtype as follows: for each BC subtype, (1) we use the GISTIC tool [1] to process the variation (SNP, CNA, CNV) data and select the top n significant altered genes in the subtype; (2) we use the chi-square ranking method to process the gene expression (GE) data and select the top n differentially expressed genes in each subtype separately; (3) we take the common genes in (1) and (2) above as the initial candidate seed genes for the subtype; 4) Finally, we map the seed genes onto a PPI network and find the subtype's NB in a greedy manner. To find NB of each subtype, we perform a *one-against-all classification* scheme with 10-fold cross-validation; taking the given subtype's samples as the positive class and the other subtypes' samples as the negative class. Figure 1 shows our approach for finding NB, separately, for each subtype. The details of our approach follows below.

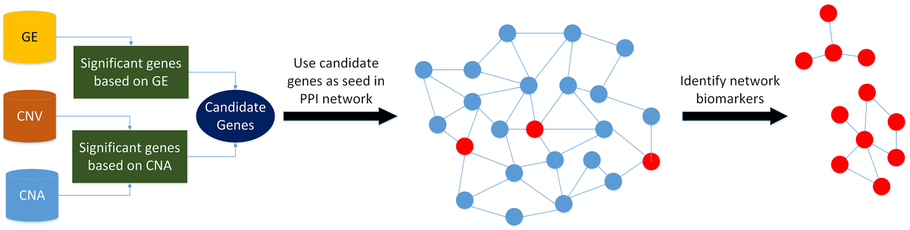
GE
Significant genes based on GE
CNV
Candidate Genes
Significant genes based on CNA
CNA
Use candidate genes as seed in PPI network
Identify network biomarkers

Fig. 1. A schematic view of the proposed framework.

Selecting the Seed Genes of a Subtype In both selection processes (steps 1 and 2 above) we overcome the class imbalance issue by using a cost-sensitive selection method that calculates different false classification costs for each class depending on their size. The seed set is the intersection the two selected gene sets for the subtype.

Finding the NB of a Subtype We map all seed genes of the subtype onto single large PPI network, to seed the search for a best separating NB. Starting from each seed node v, the search for a NB proceeds as follows. We iteratively aggregate its neighboring nodes u in a greedy manner. A random neighbor u is inserted into the current aggregate S if the correlation between the genes in $S \cup \{u\}$ and the given subtype increases; that is when $|corr(S^+ = S \cup \{u\}, subtype) - corr(S, subtype)| > \Delta$ (we set $\Delta = 0.001$ by trial-and-error). This process is repeated on the new aggregate S^+ and continues until no neighbor can be added to increase the correlation. This results in a subnetwork, S_v, obtained from a seed v. The final subtype's NB is the union of all subnetworks S_v.

3 Results

We use *accuracy* (A), F*-measure* (F) and *area under ROC curve* (AUC) as predictive performance measures of each NB. Table 2 shows the number of seed genes (SG), the number of nodes (N) and interactions (I) in each identified NB, and the phenotype correlation values (PC). Since the classes are highly imbalanced, using a more robust performance measure such as AUC provides an unbiased insights for the performance of the NB of each subtype. In Table 2, the mean AUC over all subtypes is 95.48% and 90% of the NBs have an AUC of at least 90%. This shows that our method indeed yields subtype NBs with high predictive performance.

Table 2. Sizes and predictive performances of the subtype NBs

Subtype	SG	N	I	PC	A	F	AUC
1	42	257	215	-0.7794	94.48%	0.940	0.969
2	16	242	226	-0.7567	96.79%	0.962	0.991
3	32	321	289	0.6970	87.96%	0.873	0.896
4	96	361	265	-0.6651	89.26%	0.883	0.915
5	18	195	177	-0.8195	97.39%	0.974	0.993
6	69	388	319	-0.8046	98.29%	0.982	0.997
7	16	282	266	-0.7827	92.48%	0.911	0.916
8	27	290	263	0.7729	90.87%	0.900	0.931
9	59	309	250	-0.7206	95.59%	0.949	0.968
10	75	200	125	-0.8112	96.29%	0.963	0.972

We have further analyzed some of the NBs we found. Figure 2 shows the largest component from the NB of Subtype-1 and its seed gene ARID5B (in blue). ARID5B is a well known oncogene [6] implicated in transcription regulation and involved in cell differentiation and proliferation. It has been suggested that it is involved in cancer-related signaling pathways and highly mutated in tumors.

4 Conclusion

We have introduced a novel framework for identifying the specific network biomarkers of ten breast cancer subtypes. We integrated a secondary protein interaction network

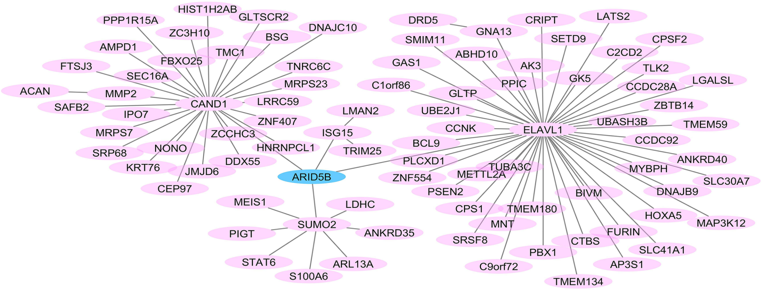

Fig. 2. The largest connected component from the NB of Subtype-1.

data with the primary tumor data consisting of gene variation and expression information to identify network biomarkers that best distinguish the subtypes. The identified network biomarkers yield high AUC results, and hence, will allow breast cancer researchers to gain additional insight into the molecular mechanisms driving each breast cancer subtype.

References

1. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al.: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences 104(50), 20007–20012 (2007)
2. Ceol, A., Aryamontri, A.C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., Cesareni, G.: Mint, the molecular interaction database: 2009 update. Nucleic acids research p. gkp983 (2009)
3. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486(7403), 346–352 (2012)
4. DeSantis, C., Ma, J., Bryan, L., Jemal, A.: Breast cancer statistics, 2013. CA: a cancer journal for clinicians 64(1), 52–62 (2014)
5. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al.: The intact molecular interaction database in 2012. Nucleic acids research p. gkr1088 (2011)
6. Lin, C., Song, W., Bi, X., Zhao, J., Huang, Z., Li, Z., Zhou, J., Cai, J., Zhao, H.: Recent advances in the arid family: focusing on roles in human cancer. OncoTargets and therapy 7, 315 (2014)
7. Liu, X., Liu, Z.P., Zhao, X.M., Chen, L.: Identifying disease genes and module biomarkers by differential interactions. Journal of the American Medical Informatics Association 19(2), 241–248 (2012)
8. Prasad, T.K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al.: Human protein reference database2009 update. Nucleic acids research 37(suppl 1), D767–D772 (2009)
9. Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al.: The biogrid interaction database: 2011 update. Nucleic acids research 39(suppl 1), D698–D704 (2011)