

Integrated Data Mining for Identification of Biomarker Candidates

Jingchun Chen^{1*}, Paul Shannon¹, Julian Watts¹, Ruedi Aebersold^{1,2}

1. Institute for Systems Biology, Seattle, Washington

2. Institute of Molecular Systems Biology, Zurich, Switzerland

*Email: jchen@systemsbiology.org

Methods and Results

One of the key components of successful cancer intervention is detecting tumor malignancy as early as possible. Recently, targeted mass-spectrometry-based proteomic analyses, such as Multiple Reaction Monitoring (MRM), allow for accurate protein quantification in biological samples, and therefore holds great potential for finding new molecular biomarkers for early cancer detection. Since such targeted analyses require genes of interest at the start, we report here an approach for the identification of breast cancer candidate marker genes using an integrated bioinformatic approach (Figure 1, panel A).

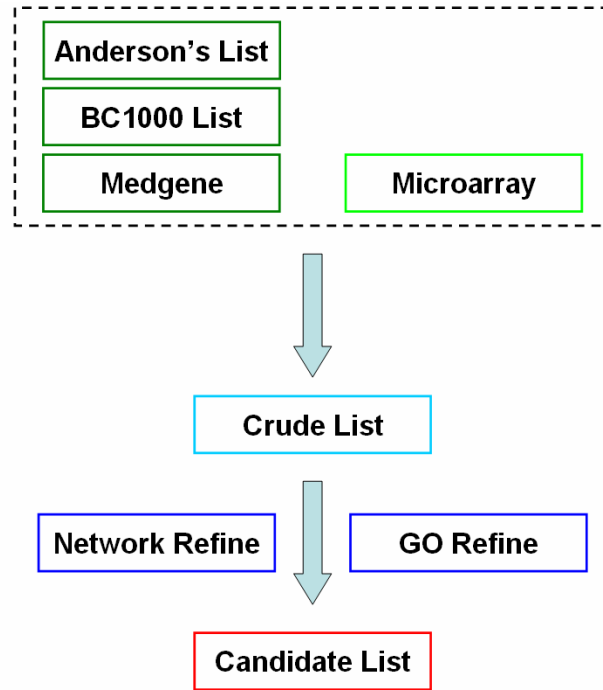
We first explored the rich information contained in numerous studies already published. Polanski *et al.* identified 1261 proteins that show differential expression in human cancer through literature searches [1]. Witt *et al.* reported over 1000 genes that are relevant to breast cancer [2]. In addition, we also searched Medgene, a database of gene-disease associations based upon literature citation [3], and retrieved about 2200 genes with varying levels of association with breast cancer.

Microarray profiling of normal and tumor tissues provides another rich source of data that can be mined for differentially expressed genes. We manually queried the Gene Expression Omnibus (GEO) database at NCBI [4], finding 7 studies that compared normal breast tissues and breast tumors of various types/stages. We identified genes that were either up- or down-regulated in tumor cells, scoring each gene with the number of differential expression evidences. For the purpose of validation, we then selected the top-ranked genes from the list and built a protein-protein interaction network. As shown in Figure 2, the network contains a large cluster, indicating that these genes indeed are functionally related ($p < 10^{-5}$).

Next we scaled the scores in the gene lists derived from literature and microarray analysis so that they are comparable, and then combined the lists by averaging each gene score. This crude candidate list was then refined in the context of protein interactions and biological processes. A gene that was not in the crude list, but interacted with genes that were on the list, was added to the list and given a score of the mean scores of its interacting partners. We next utilized gene annotations from the Gene Ontology (GO) database [5], which captures functional relationships of genes at a higher level. Genes that belonged to the same biological process were given score of $-\log(p)$. After combining these refinement scores with the initial crude scores, the average score was taken to represent the biological relevance of a gene as a potential marker for breast cancer. As shown in Figure 1, panel B, a small set of genes have high scores, while most have much lower scores. This result suggests that the scoring and integration schemes improved the signal to noise ratio. Those genes with high scores may thus represent better candidates for targeted MRM analyses.

Finally, in order to apply this data mining approach to candidate marker identification for other diseases, we are building a Java package to automate the process.

A



B

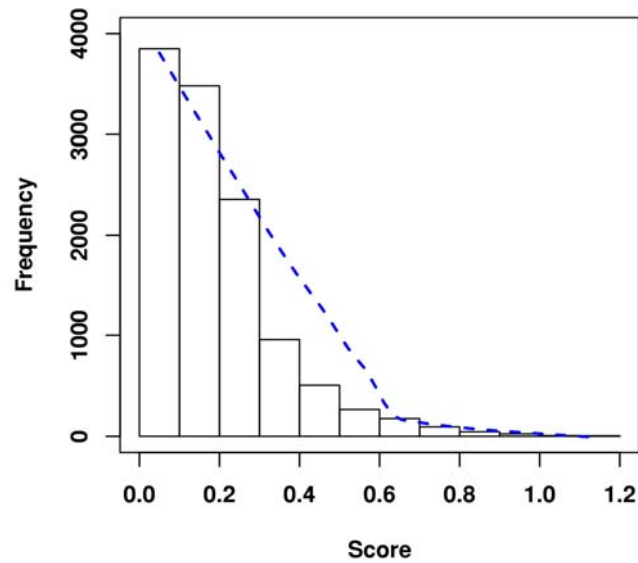


Figure 1: Integrated data mining for identifying biomarker candidates. (A) Schematic of the data analysis and integration strategy. (B) Histogram of the scores of all candidate genes derived from all data sources. The blue dashed line represents a curve fitting of second order for the score frequency, which suggests that 0.6 is a good value for the threshold to be used to select top-ranked genes.

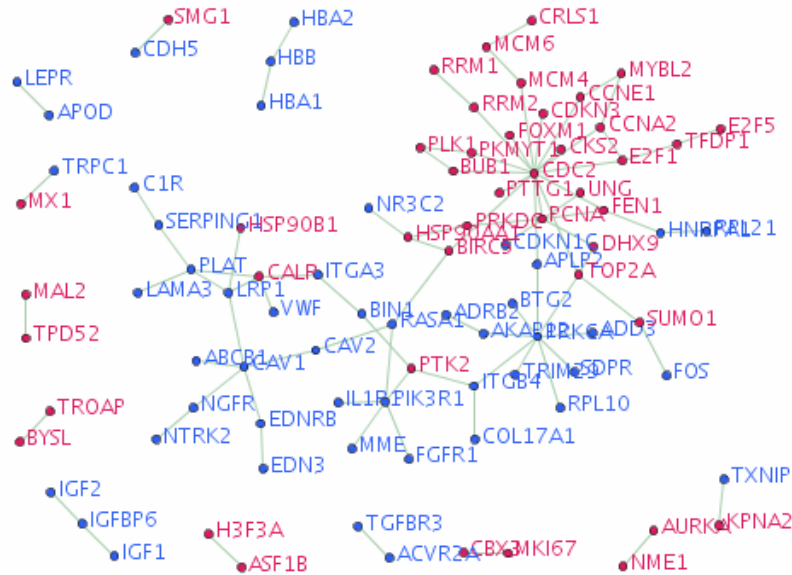


Figure 2: A protein-protein interaction network of top-ranked genes derived from differential expression analysis of microarray datasets from GEO. Up-regulated genes are colored red, and down-regulated genes are colored blue.

Acknowledgments

This work was supported in part with federal funds from the NHLBI, National Institutes of Health, under Contract N01-HV-28179 (to R.A.) and by the Entertainment Industry Foundation and its Women’s Cancer Research Fund (to J.W.).

References

1. Polanski M & Anderson NL (2006). Biomarker Insights, 2: 1-48.
2. Witt AE, Hines LM, Collins NL, *et al* (2006). Journal of Proteome Research, 5: 599-610.
3. Hu Y, Hines LM, Weng H, *et al* (2003). Journal of Proteome Research, 2(4): 405-412.
4. Barrett T, Troup DB, Wilhite SE, *et al* (2007). Nucleic Acids Research, 35: D760-D765.
5. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000). Nature Genet. 25: 25-29.