The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

The network biomarker discovery in prostate cancer from both genomics and proteomics levels

Guangxu Jin[1] Xiaobo Zhou[1,*] Kemi Cui[1]
Stephen T.C. Wong[1]

[1]Center for Bioinformatics and Biotechnology,
The Methodist Hospital Research Institute and Cornell University, Houston, TX 77030
[*]XZhou@tmhs.org

Abstract Both mass spectrometry (MS) and microarray technologies are very promising for discovery of new biomarkers for clinical diagnosis. In order to identify the high-confidence biomarkers from expression datasets, we proposed a new pipeline for biomarker discovery, in which the disease information of proteins/genes, different levels of expression profiles (microarray datasets and proteomics datasets), and interactions between proteins have been integrated. In our analysis, we first identified 474 molecules (genes and proteins) related to prostate cancer from Ingenuity software and built up a prostate-cancer-related network (PCRN) by searching the interactions among these found proteins. Based on the PCRN, the network biomarkers are discovered from multiple expression profiles composed by eight microarray datasets and one proteomics dataset. Through combining expression profiles of different levels and the protein information, we derived the network biomarkers with protein-protein interactions, which display high-performances in patient classification of prostate cancer.

Keywords Network biomarker; Mass spectrometry; Microarray; Prostate cancer

1 Introduction

The discovery, identification, and validation of differently expressed proteins, or biomarkers, in clinical proteomics are very helpful and promising for early diagnosis and treatment of many diseases. One of the novel strategies employed for discovering new biomarkers is Surface enhanced laser desorption/ionization time of flight mass spectrometry (SELDI-TOF-MS), in which the proteins contained in plasma or blood serum will enter into biological analysis [1]. SELDI-TOF-MS has been widely used to detect biomarkers in prostate [2, 3], ovarian [4], bladder [5], and breast cancers [6, 7]. Systematic proteomic studies to discover biomarkers are imperative since proteins perform the main cellular functions essential to signal transduction that lead to cell growth, differentiation, proliferation and death.

Despite a large interest and investment in this area, only a few new proteomics biomarkers were successfully used in clinical application. According to the report of the US Food

and Drug Administration (FDA), the proteomics biomarker rate of introduction is falling every year and the rate of introduction of new protein analytes approved by FDA has fallen to one per year on an average since 1998 [8, 9]. The major challenge of discovering biomarkers from proteomics is the inconsistent and irreproducible biomarker candidates. Bioinformatics algorithms for biomarker candidate discovery include baseline removal, normalization, denoising/smoothing, peak detection, peak alignment, feature selection, classification for biomarker candidates, protein/peptide identification. Many different ways available to perform each step in the process of biomarker discovery, it often results in diverse outcomes from the use of different combinations of the algorithms. Obviously, it is not feasible to overcome the challenges in biomarker discovery only depending on the improvement of these bioinformatics algorithms. Therefore, a better choice is to combine more known and reliable protein-related information into proteomics biomarker discovery process.

Here, we combined disease information of proteins [10], protein-protein interactions [11], and multiple-microarray expression profiles [12] into proteomics biomarker discovery. Actually, we made a tradeoff between protein knowledge and data noises in MS. To make sure that the found proteins are really useful for prostate cancer, we filtered out PCRN, in which each protein has been identified as related to prostate cancer by publications and involved in at least one protein-protein interaction in HPRD database. Then we identified differentially expressed proteins from the proteins in PCRN. There are two advantages in such a biomarker discovery process. First, the tradeoff between protein information and MS data noise can make the found biomarkers more confident and accurate than identification of biomarkers directly from MS data. That is due to the fact that PCRN is composed of those proteins already reported by some big journals or annotations of some databases, such as Uniprot, HPRD, KEGG, or others. Next, the biomarkers found by our pipeline are differently expressed not only at the proteomics level but also at the genomic level and thereby they are more powerful in the patient classification. In this paper, PCRN with 131 protein and 310 interactions was identified. Based on the identified network, we proposed a new type of biomarkers, i.e. network biomarkers with protein-protein interactions, which have relatively high classification accuracy for prostate cancer patients.

2 Results

2.1 PCRN construction

Most biomarkers nowadays are discovered from distinct molecular levels, such as microarray and proteomics expression profiles [13-19]. But one big problem with them is their inconsistency and irreproducibility from the discovery process. Since the biomarkers aimed to indicate the specific case of a disease, such as prostate cancer, they should somehow have a close relationship with the studied disease. Undoubtedly, the known disease information of a protein should be essential to biomarker discovery. However, few previous studies have considered the importance of the information to biomarker discovery. Our biomarker discovery pipeline considered such important disease information and applied it to reduce the occurrence probability of false-positive biomarkers caused by data noises in microarray and MS data (Figure 1). A beautiful software(IPA) can provide such disease information of proteins based on its manually integrated published results in

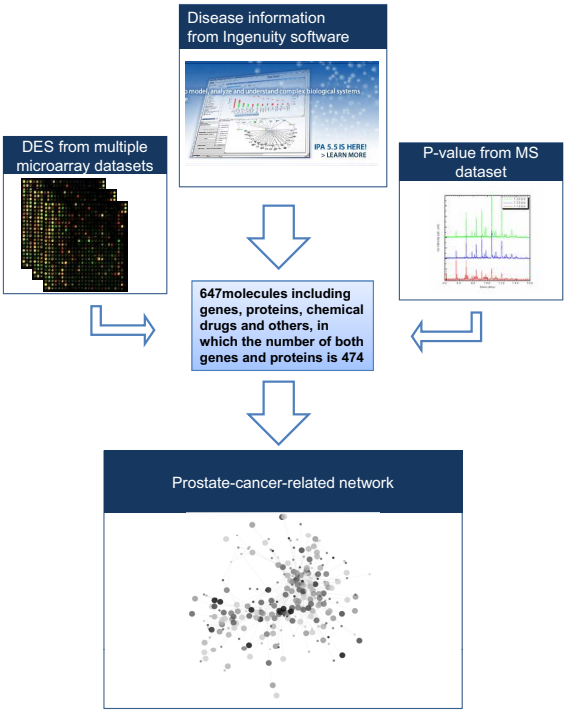

Figure 1: The flowchart for biomarker discovery from prostate cancer.

some big journals, by which about 400 proteins and genes related to prostate cancer has been found. Before combination of multiple genomics and proteomics expression profiles, we first built up a prostate-cancer related network (PCRN). To analyze and explain the important roles of PPIs, we filtered out some prostate-cancer-related proteins with at least one PPI from HPRD. The found biomarkers based on PCRN include not only the single proteins but also the network biomarkers with PPIs.

2.2 Differential expressions for the proteins in PCRN

The extra advantage of our method is assorting available and reliable protein information to reduce the errors in biomarker discovery. Combination of multiple microarray datasets with proteomics expression profiles is an essential strategy for that purpose, in which the Differential expression scores (DESs) are necessary to find those genes with highly differential expressions between normal patients and prostate patients.

Eight microarray datasets from different experiments have been considered in the process of biomarker discovery (Table 1). We computed the p-value (student t test, significance level: 0.05) for every gene for each microarray dataset so that we can find those genes with significantly low p-values for normal and prostate patients. In consideration that the data-noises in microarray datasets can bring many troubles to biomarker discovery process, we chose the genes displaying low p-values in multiple microarray datasets. The DES value can indicate to what extent a gene is expressed differentially for the pa-

tients in two conditions. If the DES of a gene is very high, it indicates the gene displays differential expressions nearly in all experiments, which implies that the gene is most likely to be expressed differentially in real situation. Thus, we can reduce the data noises for biomarker discovery by combination of multiple microarray datasets.

Table 1: The microarray datasets in Oncomine used in definition of DES

Study name	Samples
Lapointe_Prostate	Normal Prostate (41), Prostate Carcinoma (62)
Yu_Prostate	Normal Prostate (23), Prostate Carcinoma (64)
Welsh_Prostate	Normal Prostate (9), Prostate Carcinoma (25)
Dhanasekaran_ Prostate_2	Normal Adjacent Prostate (2), Normal Adult Prostate (7), Normal Pubertal Prostate (3), Prostate Cancer (25)
Vanaja_Prostate	Normal Prostate (8), Prostate Adenocarcinoma (27)
Singh_Prostate	Normal Prostate (50), Prostate Carcinoma (52)
Holzbeierlein_Prostate	Normal Prostate (4), Prostate Cancer (23)
Magee_Prostate	Normal Prostate (4), Prostate Cancer (8)

Similarly, we also computed the P-value for the peak intensities of a protein for the 81 normal patients and 168 prostate cancer patients from MS data (Materials and Methods). If the P-value is low, it implies that the protein is differentially expressed in the two types of patient from proteomics level. This step is also a necessary step for candidate biomarker discovery on MS data.

With the two levels of expression profiles, i.e. genomics and proteomics, the filtered proteins in this step have a low p-value in MS data and their corresponding genes have a high DES in multiple microarray datasets. Thus, we filtered out relatively high confidence proteins with differential expressions in both genomics and proteomics levels.

2.3 MS-based biomarker discovery (candidate single biomarkers)

MS-based biomarker discovery is to identify proteins differentially expressed in the serum or plasma of prostate cancer patients. A new and emerging technology, i.e. proteomics, has the potential to identify protein molecules in a high-throughput discovery approach in patient's serum. For a protein, its mass in the mass/charge axis was first identified and then its nearest peak or mean of the masses in the window of -10Da and +10Da of its mass has been identified as one of the expression intensities for the protein. Thus, the intensity vectors for different conditions can be derived from 81 normal and 168 prostate patients.

To elucidate the significantly distinct expressions of a protein between control and prostate cancer patients, we still adopted p-value from student t test to discover biomarkers. If the intensity vectors for a protein are affected by the data noises significantly, the P-value to evaluate different expressions of the protein will not be significantly low and the peptide will not provide evidence of its protein as a candidate biomarker discovered from MS. On the other hand, the protein with relatively low P-value implies that its intensities should not be greatly disturbed by the data noises because they are differently expressed in control and disease patients, and thereby be considered as a candidate biomarker.

2.4 Network biomarker identification on PCRN

The protein-protein interaction information in PCRN was not considered in the identification process of candidate single biomarkers for MS data. The interactions between proteins are important for many biological functions. Due to the essential roles of protein interactions in biological processes, we integrated the protein-protein interaction information into the biomarker discovery process. We revealed a new type of biomarkers, called as network biomarkers, composed of a set of proteins with the interactions among them.

Network biomarkers considered in our analysis can be divided into three types, single biomarker without any protein-protein interaction, pair-biomarker with two proteins and one protein-protein interaction, triple-biomarker with three proteins and three protein-protein interactions, square-biomarker with four proteins and four protein-protein interactions.

Let P_i be the protein involved in a network biomarker, p_i be the P-value for P_i. And also let $\mathbf{I}_i$ be the intensity vector of protein P_i for not only 81 control patients but 168 prostate cancer patients, then the combined intensity vector $\mathbf{I}_{network}$ for the pair biomarker is

$$\mathbf{I}_{network} = \sum_{i=1}^{N} \frac{\frac{1}{p_i}}{\sum_{j=1}^{N} \frac{1}{p_j}} \mathbf{I}_i$$

Classification based on SVM was applied to identify network biomarkers based on their classification performances in testing sets. Here, a 5-fold cross validation in SVM was used to classify patients in control and prostate cancer patients. The training set for each split included 4/5 of the cases, while 1/5 of the samples were used as the test set and were not involved in training. First, different number, 1 or 2 or 3, of same type of network biomarkers was put into SVM. By their performance, we can easily identify the best ones for patient classification. We found that the best performances for single biomarkers, P35222, P55210, and P15941, is 85.14% , the best performance for pair biomarkers, P00734-P10451, Q14790-p12830, and P12830-P35222 is 85.94%, the best performance for triple biomarkers, P10451-P24593-P00734, P00747-P10344-P17936, and P04004-P24593-P01344 is 80.72%, and the best performance for square biomarkers, O15393-P55210-Q05513-P55211, P00749-P00747-P17936-P00734, and Q14790-Q05513-P55210-O15379 is 79.91% (Table 2).

Table 2: The classification accuracies for different network biomarkers.

#	Single	Pair	Triple	Square
1	77.10%	77.10%	76.71%	76.71%
	P05109	P15109-P62937	P10451-P24593-P00734	O15392-P55210-Q05513-P55211
2	81.53%	81.12%	80.72%	79.12%
	P10145	P14780-P10145	P35222-P12830-Q14790	O15392-P55210-Q05513-P55211
	P55210	P55210-O15519	P00747-P01344-P17936	P24593-P01344-P17936-P00734
3	85.14%	85.94%	80.72%	79.92%
	P35222	P00734-P10451	P10451-P24593-P00734	O15393-P55210-Q05513-P55211
	P55210	Q14790-P12830	P00747-P01344-P17936	P00749-P00747-P17936-P00734
	P15941	P12830-P35222	P04004-P24593-P01344	Q14790-Q05513-P55210-O15379

Next, we put the single and pair biomarkers into SVM and found that the combination of these two relatively high-confidence biomarker can derive the same high classification accuracy as pair biomarkers, i.e. 85.94%. The multi-type biomarker is shown in Table 3.

Table 3: The classification accuracies for the combination of single and pair biomarkers.

Combination	2 Single and 1 Pair	1 Single and 2 Pair
Accuracy	85.94%	85.94%
Biomarkers	P15941	P35222
	P11362	Q14970-P12830
	Q14970-P12830	P00734-P10451

3 Discussion

Our proposed pipeline for biomarker discovery is composed of integration of disease information of proteins, combination of multiple microarray and proteomics expression profiles, and PCRN construction. Instead of focusing on improvement the existing bioinformatics methods in proteomics biomarker discovery, we assorted the available and reliable protein information to derive the high confidence biomarkers from MS data. In this manner, we can easily overcome the difficulties in discovering proteomics biomarkers from MS data and identify the high confidence network biomarkers.

Most previous biomarker discovery works focused on either proteomics level or genomics level and few of them studied biomarkers from both proteomics and genomics levels. In our paper, we proposed such a method based on not only proteomics expression profile, i.e. MS data, but also multiple genomics expression profiles, i.e. Microarray data. The biomarkers found in this way can consistently display their differential expressions not only in mRNA expressions but also in protein expressions. Therefore, such a method is very promising for discovering biomarkers from expression profiles with relatively high data noises.

How to overcome the data noises in high-throughput datasets, such as microarray and MS, is a big problem for biomarker discovery. In this paper, we used a voting method for biomarker discovery, that is, the found biomarkers should be simultaneously supported by multiple microarray datasets, proteomics dataset, the disease information from Ingenuity, PPIs of HPRD, and high classification accuracies in SVM. The high accuracies of found biomarkers in patient classification indicate that our method has the power to derive the high-confidence biomarkers.

Acknowledges

This research is funded by the Bioinformatics Core Research Grant at The Methodist Research Institute, Cornell University. Dr. Zhou is partially funded by The Methodist Hospital Scholarship Award. He and Dr. Wong are also partially funded by NIH grants R01LM08696, R01LM009161, and R01AG028928. The authors would like to thank Prof. Xiang-Sun Zhang and Prof. Luonan Chen for helpful discussions and suggestions.

References

[1] Issaq, H. J., Veenstra, T. D., Conrads, T. P., Felschow, D., The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem Biophys Res Commun 2002, 292, 587-592.

[2] Malik, G., Rojahn, E., Ward, M. D., Gretzer, M. B., et al., SELDI protein profiling of dunning R-3327 derived cell lines: identification of molecular markers of prostate cancer progression. Prostate 2007, 67, 1565-1575.

[3] Paweletz, C. P., Liotta, L. A., Petricoin, E. F., 3rd, New technologies for biomarker analysis of prostate cancer progression: Laser capture microdissection and tissue proteomics. Urology 2001, 57, 160-163.

[4] Oh, J. H., Nandi, A., Gurnani, P., Knowles, L., et al., Proteomic biomarker identification for diagnosis of early relapse in ovarian cancer. J Bioinform Comput Biol 2006, 4, 1159-1179.

[5] Vlahou, A., Schellhammer, P. F., Mendrinos, S., Patel, K., et al., Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001, 158, 1491-1502.

[6] Wulfkuhle, J. D., McLean, K. C., Paweletz, C. P., Sgroi, D. C., et al., New approaches to proteomic analysis of breast cancer. Proteomics 2001, 1, 1205-1215.

[7] Pal, S. K., Pegram, M., HER2 targeted therapy in breast cancer...beyond Herceptin. Rev Endocr Metab Disord 2007, 8, 269-277.

[8] Anderson, N. L., Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 2002, 1, 845-867.

[9] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., et al., The human plasma proteome: a nonredundant list developed by combination of four separate sources. Mol Cell Proteomics 2004, 3, 311-326.

[10] Cornish, E. J., Hurtgen, B. J., McInnerney, K., Burritt, N. L., et al., Reduced nicotinamide adenine dinucleotide phosphate oxidase-independent resistance to Aspergillus fumigatus in alveolar macrophages. J Immunol 2008, 180, 6854-6867.

[11] Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., et al., Human protein reference database–2006 update. Nucleic Acids Res 2006, 34, D411-414.

[12] Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., et al., ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004, 6, 1-6.

[13] Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., Ideker, T., Network-based classification of breast cancer metastasis. Mol Syst Biol 2007, 3, 140.

[14] Allantaz, F., Chaussabel, D., Banchereau, J., Pascual, V., Microarray-based identification of novel biomarkers in IL-1-mediated diseases. Curr Opin Immunol 2007, 19, 623-632.

[15] Forrest, M. S., Lan, Q., Hubbard, A. E., Zhang, L., et al., Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. Environ Health Perspect 2005, 113, 801-807.

[16] Izuhara, K., Saito, H., Microarray-based identification of novel biomarkers in asthma. Allergol Int 2006, 55, 361-367.

[17] Dihazi, H., Muller, G. A., Urinary proteomics: a tool to discover biomarkers of kidney diseases. Expert Rev Proteomics 2007, 4, 39-50.

[18] Mayr, M., Zhang, J., Greene, A. S., Gutterman, D., et al., Proteomics-based development of biomarkers in cardiovascular disease: mechanistic, clinical, and therapeutic insights. Mol Cell Proteomics 2006, 5, 1853-1864.

[19] Zinkin, N. T., Grall, F., Bhaskar, K., Otu, H. H., et al., Serum proteomics and biomarkers in hepatocellular carcinoma and chronic liver disease. Clin Cancer Res 2008, 14, 470-477.

[20] Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., et al., Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 2004, 101, 811-816.

[21] Yu, Y. P., Landsittel, D., Jing, L., Nelson, J., et al., Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. J Clin Oncol 2004, 22, 2790-2799.

[22] Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., et al., Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res 2001, 61, 5974-5978.

[23] Dhanasekaran, S. M., Dash, A., Yu, J., Maine, I. P., et al., Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty. FASEB J 2005, 19, 243-245.

[24] Vanaja, D. K., Cheville, J. C., Iturria, S. J., Young, C. Y., Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. Cancer Res 2003, 63, 3877-3882.

[25] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., et al., Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002, 1, 203-209.

[26] Holzbeierlein, J., Lal, P., LaTulippe, E., Smith, A., et al., Gene expression analysis of human prostate carcinoma during hormonal therapy identifies androgen-responsive genes and mechanisms of therapy resistance. Am J Pathol 2004, 164, 217-227.

[27] Magee, J. A., Araki, T., Patil, S., Ehrig, T., et al., Expression profiling reveals hepsin overexpression in prostate cancer. Cancer Res 2001, 61, 5692-5696.

[28] Adam, B. L., Qu, Y., Davis, J. W., Ward, M. D., et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002, 62, 3609-3614.

[29] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000, 16, 906-914.

[30] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., et al., Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000, 97, 262-267.

[31] Jin, G., Zhou, X., Wang, H., Zhao, H., et al., The Knowledge-integrated network biomarkers discovery for Major Adverse Cardiec Events. J Proteome Res 2008, 7(9): 4013-4021..