Computational Biology and Chemistry 49 (2014) 71–78

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Improving the prediction of chemotherapeutic sensitivity of tumors in breast cancer via optimizing the selection of candidate genes

Lina Jiang[a], Liqiu Huang[a], Qifan Kuang[a], Juan Zhang[a], Menglong Li[a], Zhining Wen[a,*], Li He[b,**]

[a] College of Chemistry, Sichuan University, Chengdu 610064, PR China
[b] Biogas Institute of Ministry of Agriculture, Chengdu 610041, PR China

ARTICLE INFO

Article history:
Received 25 September 2013
Received in revised form
14 December 2013
Accepted 17 December 2013
Available online 1 January 2014

Keywords:
Cancer outcome prediction
Gene expression profiling
Gene prioritization
Support vector machine
Breast cancer

ABSTRACT

Estrogen receptor status and the pathologic response to preoperative chemotherapy are two important indicators of chemotherapeutic sensitivity of tumors in breast cancer, which are used to guide the selection of specific regimens for patients. Microarray-based gene expression profiling, which is successfully applied to the discovery of tumor biomarkers and the prediction of drug response, was suggested to predict the cancer outcomes using the gene signatures differentially expressed between two clinical states. However, many false positive genes unrelated to the phenotypic differences will be involved in the lists of differentially expressed genes (DEGs) when only using the statistical methods for gene selection, e.g. Student's *t* test, and subsequently affect the performance of the predictive models. For the purpose of improving the prediction of clinical outcomes, we optimized the selection of DEGs by using a combined strategy, for which the DEGs were firstly identified by the statistical methods, and then filtered by a similarity profiling approach that used for candidate gene prioritization. In our study, we firstly verified the molecular functions of the DEGs identified by the combined strategy with the gene expression data generated in the microarray experiments of *Si-Wu-Tang*, which is a popular formula in traditional Chinese medicine. The results showed that, for *Si-Wu-Tang* experimental data set, the cancer-related signaling pathways were significantly enriched by gene set enrichment analysis when using the DEG lists generated by the combined strategy, confirming the potentially cancer-preventive effect of *Si-Wu-Tang*. To verify the performance of the predictive models in clinical application, we used the combined strategy to select the DEGs as features from the gene expression data of the clinical samples, which were collected from the breast cancer patients, and constructed models to predict the chemotherapeutic sensitivity of tumors in breast cancer. After refining the DEG lists by a similarity profiling approach, the Matthew's correlation coefficients of predicting estrogen receptor status and the pathologic response to preoperative chemotherapy with the DEGs selected by the fold change ranking were 0.770 and 0.428, respectively, and were 0.748 and 0.373 with the DEGs selected by SAM, respectively, which were generally higher than those achieved with unrefined DEG lists and those achieved by the candidate models in the second phase of Microarray Quality Control project (0.732 and 0.301, respectively). Our results demonstrated that the strategy of integrating the statistical methods with the gene prioritization methods based on similarity profiling was a powerful tool for DEG selection, which effectively improved the performance of prediction models in clinical applications and can guide the personalized chemotherapy better.

1. Introduction

In breast cancer treatment, the chemotherapeutic sensitivity of tumor is crucial for guiding the selection of the most effective chemotherapy for a specific patient (Early Breast Cancer Trialists Collaborative Group, 1998). The estrogen receptor status is considered as one of the important indicators of chemotherapeutic sensitivity in diagnostic test (Bast et al., 2001; Ross et al., 2003; Paik et al., 2006), for which the estrogen receptor positive (ER positive) indicates that the growth of the tumor is caused by estrogen and should be effectively suppressed by hormone suppression treatments, while the estrogen receptor negative (ER negative) indicates bad response to hormone suppression treatments. Differ from the estrogen receptor status, the pathologic response to preoperative chemotherapy, a surrogate endpoint in neoadjuvant treatment, can directly evaluate the tumor response to the chemotherapy for the individual patient (Fisher et al., 1998; Kuerer et al., 1999). The

* Corresponding author. Tel.: +86 28 85412138; fax: +86 28 85412356.
** Corresponding author. Tel.: +86 28 85230702.
E-mail addresses: w_zhining@163.com (Z. Wen), heliscu@gmail.com (L. He).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.12.002

pathologic complete response (pCR) indicates the well postoperative recovery and long-term survival for the patients, while residual disease (RD) means the high risk of recurrence. Consequently, it is necessary to accurately obtain the information about the estrogen receptor status and the tumor response to preoperative chemotherapy for the purpose of determining the best regimen and achieving the best clinical outcomes in breast cancer treatments.

The gene expression profiling generated by the DNA microarrays is highly associated with the phenotypic differences between the clinical samples and has been suggested to investigate the molecular functions, impacted signaling pathways and cellular behaviors for better understanding of the molecular mechanisms of breast cancer (Perou et al., 2000; Weigelt et al., 2005a,b; Sotiriou et al., 2006; Pusztai et al., 2007; Liu et al., 2008; Wirapati et al., 2008; Loi et al., 2009; Daves et al., 2011; Mohammadi et al., 2011). The differentially expressed genes (DEGs) are routinely identified from the microarray-based gene expression profiles and used as the predictors for the prediction of chemotherapeutic response and clinical outcomes in breast cancer trials (van't Veer et al., 2002; Ma et al., 2004; Goetz et al., 2006; Hess et al., 2006; Nuyten and van de Vijver, 2006–2007; Cronin et al., 2007; Ross, 2009; Lee et al., 2010; Yau et al., 2010; Fan et al., 2011; Patsialou et al., 2012; Hallett et al., 2012). The Microarray Quality Control (MAQC) project systematically evaluated the reproducibility and reliability of the DEG lists generated by different microarray platforms (Shi et al., 2006) and the reliability of models for predicting the preclinical and clinical endpoints with the selected DEGs, providing the best practices on the development and validation of microarray-based predictive models (Shi et al., 2010). However, most of the statistical methods and feature selection methods used in DEG identification can only discriminate between DEG and non-DEG according to the differences of gene expression levels between two biological states, but do not ensure the biological relevance of a selected DEG to the phenotypic differences, resulting in many false positive genes unrelated to the phenotypic differences involving in the DEG lists.

For the sake of screening out the false positive genes and improving the performance of the predictive models, we suggested a combined strategy, for which the DEGs were firstly identified by the statistical methods, and then filtered by a similarity profiling approach that used for candidate gene prioritization. The gene prioritization methods based on similarity profiling ranked the candidate genes according to their similarity scores, which were calculated by considering the similarities of the candidate genes to a set of known seed genes in a specific disease or biological process, and identified the most promising genes with the maximal relevance to that disease or biological process (van Driel et al., 2003; Aerts et al., 2006; Tranchevent et al., 2008; Seelow et al., 2008; Chen et al., 2009; Fontaine et al., 2011; Britto et al., 2012; Moreau and Tranchevent, 2012). As a complementary method to DEG identification, the similarity profiling approach should well select the genes actually related to phenotypic differences from the DEG list generated by statistical methods. In our study, the breast cancer-related gene sets were collected by the similarity profiling approach from the DEG lists generated by two statistical methods, fold change ranking combined with a non-stringent p value cutoff and significant analysis of microarrays (SAM). In order to answer the two questions: (a) whether the gene sets identified by the combined strategy were still significantly correlated with the phenotypic status and (b) whether the DEGs identified by the combined strategy can improve the performance of the predictive models in clinical research, we conducted a comparative study, which included two components. The one is the verification of the molecular functions of the gene sets identified by the combined strategy using the gene expression data generated in the microarray experiments of *Si-Wu-Tang* (Wen et al., 2011). In this part of the work, we examined whether the gene sets identified by the combined strategy from the microarray data generated from the MCF-7 cell lines with and without treatment (*Si-Wu-Tang*) were still significantly correlated with cancer. The other is the verification of the performance of the predictive models. The gene expression data of the clinical samples from 230 breast cancer patients were collected and separated into training set (130 samples) and validation set (100 samples). We selected the DEGs as features by using the combined strategy and constructed the predictive models to predict the chemotherapeutic sensitivity of tumor by using the training set. The predictive models were validated by the validation set. Our results demonstrated that the strategy of integrating the statistical methods with the similarity profiling-based gene prioritization methods for DEG identification can effectively enhance the reliability of the predictive models and improve the prediction results of clinical outcomes in breast cancer.

2. Materials and methods

2.1. Data set

The data set was comprised of two subsets generated from Michigan Cancer Foundation-7 (MCF-7) cell lines before and after treatment by *Si-Wu-Tang* (series accession number: GSE23610) (Wen et al., 2011) and the clinical samples of 230 breast cancer patients (series accession number: GSE16716) (Hess et al., 2006; Shi et al., 2006; Popovici et al., 2010). All microarray data (Series MATRIX files) were downloaded through the National Center for Biotechnology Information's Gene Expression Omnibus. In the previous subset, the gene expression data were generated by using Affymetrix Human Genome U133Plus2 microarrays, which included 54,675 probesets. Only the gene expression data of MCF-7 cell lines before and after treatment with high concentration (2.56 mg/ml) of *Si-Wu-Tang* (SH) were used to examine the molecular functions of gene sets identified by our combined strategy. For the latter one, the gene expression data were generated by using Affymetrix Human Genome U133A microarrays, which included 22,283 probesets. According to the data analysis protocol in MAQC-II project, the gene expression data generated from 130 out of 230 clinical samples of breast cancer patients were assigned as training set and the rest were used as independent test set. The two prediction endpoints, the estrogen receptor status and the pathologic response to preoperative chemotherapy, were determined by immunohistochemistry and histopathologic study, respectively. A case was considered as ER positive where $\geq$10% of tumor cells stained positive for estrogen receptor with immunohistochemistry. A pCR was defined as no residual invasive cancer in the breast or lymph nodes, or the residual in situ carcinoma without invasive component (Hess et al., 2006).

2.2. Probesets mapping

Microarray experiments detected the fluorescence intensity on the individual probeset to infer a transcript abundance level of a specific gene. A gene may be detected by multiple probesets on the arrays. Before identifying DEGs, the multiple probesets were mapped to a unique HOGO gene symbol by using the probeset with the highest fold change value between two groups of samples. Accordingly, for the subsets of the *Si-Wu-Tang* experimental data set and the expression data of breast cancer samples, 54,675 probesets were mapped to 15,961 unique genes and 22,283 probesets were mapped to 11,285 unique genes, respectively.