Journal of Theoretical Biology 317 (2013) 62–70

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Prediction of core cancer genes using multi-task classification framework

Shan Gao [a,1], Shuo Xu [b,1], Yaping Fang [a], Jianwen Fang [a,*]

[a] Applied Bioinformatics Laboratory, The University of Kansas, 2034 Becker Drive, Lawrence, KS 66047, USA
[b] Institute of Scientific and Technical Information of China, No. 15 Fuxing Road, Haidian District, Beijing 100038, PR China

HIGHLIGHTS

- A strategy based on multi-tasking learning is proposed to predict core cancer genes.
- Biological significance of these genes is evaluated using systems biology analyses.
- The strategy can be used as a general method to find important features.

ARTICLE INFO

Article history:
Received 26 September 2011
Received in revised form
30 July 2012
Accepted 18 September 2012
Available online 3 October 2012

Keywords:
Multi-task learning
Classification
Core cancer genes
Gene differential expression
Microarray data

ABSTRACT

Cancer is deemed as a highly heterogeneous disease specific to cell type and tissue origin. All cancers, however, share a common pathogenesis. Therefore, it is widely believed that cancers may share common mechanisms. In this study, we introduce a novel strategy based on multi-tasking learning methods to predict core cancer genes shared by multiple cancers in the hope of elucidating common cancer mechanisms. Our strategy uses two multi-tasking learning algorithms, one for feature selection and the other for validation of selected features. The combined use of two methods results in more robust classifiers and reliable selected features. The top 73 significant features, mapped to 72 genes, are selected as core cancer genes. The effectiveness of the 73 features is further demonstrated in a blind test conducted on an independent test data. The biological significance of these genes is evaluated using systems biology analyses. Extensive functional, pathway and network analysis confirms findings in previous studies and brings new insights into common cancer mechanisms. Our strategy can be used as a general method to find important genes from large gene expression datasets on the genomic level. The selected genes can be used to predict cancers.

1. Introduction

Cancer comprises more than 100 distinct diseases specific to cell type and tissue origin (Stratton et al., 2009). All these diseases, nevertheless, share key characteristics such as uncontrolled growth and spread of abnormal cells (Lauffenburger and Kreeger, 2010). Therefore it is widely believed that all cancers share a common pathogenesis (Stratton et al., 2009). Elucidating common cancer mechanisms will certainly enhance our ability to devise effective therapeutics (Khalil and Hill, 2005) against the disease responsible for one in eight deaths worldwide (Stratton et al., 2009).

Over the past several years, a few attempts have been made to identify the core cancer genes, or the meta-signatures across a wide range of cancer types by analyzing genome-wide gene expression profiles from multiple microarray data sets (Lu et al., 2007; Rhodes et al., 2004) (Segal et al., 2004), in the hope of discovering common cancer mechanisms. These studies are part of an emerging biological domain termed as cancer systems biology (Lauffenburger and Kreeger, 2010). Computationally, the overall strategy in previous studies is to break the problem into a number of sub problems, each of which is corresponding to a learning task on a microarray data set for a specific type of cancer. After solving them separately by independent and pairwise univariate analysis (e.g. *t*-test), the results are then combined to identify the intersection of significant differentially-expressed genes. Thus, these approaches are single-task learning (STL) methods in nature and do not consider the correlations between the sub problems (single tasks, STs). Such approaches can only find the overlap of caner type-specific genes, rather than cancer type-independent genes of multiple cancers (Dawany et al., 2011). To overcome the drawbacks of these methods, we propose a novel approach within the multi-task learning (MTL) framework to find the core cancer genes by simultaneously solving those STs.

Using a shared representation, MTL learns all participated STs of a problem simultaneously by a global optimization approach

* Corresponding author. Tel.: +1 785 864 3349.
E-mail address: jwfang@ku.edu (J. Fang).
[1] These authors contributed equally to this work.

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.09.027

based on an intuitive idea: the common knowledge shared by related STs in a specific domain helps improving the performance (Caruana, 1993). It has been empirically and theoretically demonstrated that MTL can improve learning performance, compared to learning STs separately (Argyriou et al., 2006). In addition, MTL can be used to find the common knowledge and perform feature selection to identify significant features shared by member STs. Although MTL is very promising, it had not been applied to study biological problems until very recently. For example, Zhang et al. (2010) used MTL for gene expression analysis and Xu et al. (2010) applied MLT in the prediction of subcellular location of proteins. It has also been used in the prediction of siRNA efficacy (Liu et al., 2010).

In this study, we attempt to discover core cancer genes using a novel approach within the MLT framework. Our basic idea is that the most significant features (genes) in discriminating normal samples against cancer samples of various cancer types simultaneously using the global optimization approach may reflect essential characteristic of cancers and provide key information for finding common mechanisms more effectively than conventional STL approaches. First, we compile a microarray dataset MetaCancer12 including 12 sub-datasets, each representing a ST of binary classification of cancer vs. normal samples. We then merge these 12 STs into a MLT learning to identify core cancer genes by combined use of two MTL methods, the Multi-Task LS-SVMs (MTLS-SVMs) and the Multi-Task Feature Selection (MT-Feat3). MTLS-SVMs was introduced in our previous work and MT-Feat3 is derived from a MTL framework which was originally designed for regression (Argyriou et al., 2008).

Computationally, the main novelty of this study is that our new strategy uses two MTL methods in feature selection: MT-Feat3 for selecting common features from 12 STs and MTLS-SVMs for validating the selected features. The theoretical basis of our approach is that a feature set identified by a robust feature selection algorithm should be robust to allow a different algorithm to make high-accuracy predictions, even if the selected feature set is not optimal for that algorithm (Das, 2001). Successive use of these two methods combines the advantages of filter and wrapper concepts. By doing so, the biases of the feature selection (MT-Feat3) and the model learning (MTLS-SVMs) do not interact with each other because MTLS-SVMs and MT-Feat3 implement two different STL mechanisms: data amplification and feature selection (Caruana, 1997). The data amplification mechanism of MTLS-SVMs may help lessening the potential over-fitting problem on small datasets when only one method (wrapper) is used to select and validate features (Das, 2001).

As results, we identify 73 Affymetrix probe sets, out of a total of 22,215 found in all samples, as core features of 12 cancer types. The effectiveness of 73 features is cross-validated on the training dataset and blind-tested on a large independent dataset. These 73 sets are then mapped to 72 core cancer genes. We perform systems biology analysis for these 72 genes. Our results are largely consistent with previous studies and also bring new insights into possible common mechanisms of cancers.

2. Materials and methods

2.1. Dataset construction

The gene expression dataset MetaCancer12 used in this study was compiled from the web resource ONCOMINE (http://www.oncomine.org) (Rhodes et al., 2007) in February 2011. The primary filtering criteria were set to "Differential Analysis" and "Cancer vs. Normal Analysis" to acquire the datasets fit for binary classification. The platform filtering criterion was set to "Affymetrix U133" to minimize the platform variation. From a total of 53 datasets passing the filtering steps, we selected 12 of them as the training dataset MetaCancer12 by the following additional criteria:

1) Each dataset must be specific to one cancer type which represents one single task.
2) The raw data (.CEL format) of each dataset is available online so a standard data normalization process can be applied to normalize all datasets.
3) If multiple datasets are available for any specific cancer type, the most balanced one was chosen because unbalanced data may cause predictors unreliable.
4) The largest one among datasets of the same cancer type was chosen.

Overall, MetaCancer12 covered 12 common caner types: pancreas, vulva, prostate, head–neck, leukemia, renal, lung, gastric, esophagus, skin, colon and breast. We also used the remaining 11 datasets as sub-datasets to constitute an independent test dataset MetaCancer11. The detailed description of MetaCancer12 and MetaCancer11 can be seen in the Supplementary File 1.

2.2. Data pre-process and representation

Affymetrix U133 platform includes three types: Human Genome U133 Plus 2.0 Array, Human Genome U133A 2.0 Array, and Human Genome U133A&B. These types differ from the number of probe sets presented in the chip. The shared genes of those three types of microarray are represented by 22,215 Affymetrix identifiers which are used as features to describe each sample. Cancer samples are defined as positives and normal samples as negatives. Samples in 11 sub dataset are normalized by the Robust Multi-array Average (RMA) algorithm (Irizarry et al., 2003) individually. Samples in the sub dataset of oesophagus without raw data are normalized with mean=0 and standard deviation=1 (Supplementary File 1). Finally, the ith sample is represented by N=22,215 features in such form $\vec{x}_i = (x_{i1}, x_{i2}, \ldots x_{ij} \ldots x_{iN})$.

2.3. STL methods

We build 12 independent classifiers based on Least Squares Support Vector Machines (LS-SVMs) (Suykens and Vandewalle, 1999) for 12 STs (i.e. cancer sub-datasets). LS-SVMs classifiers are also used as element classifiers in MTL (see the next section). LS-SVM performs training faster than the standard SVM without sacrificing generalization performance (van Gestel et al., 2004). A fast algorithm is important to deal with large-scale and high dimensional gene expression data for practical use. A LS-SVMs classifier is obtained by solving a restricted optimization problem as below:

$$\min_{\mathbf{w},e} \quad \frac{1}{2}\|\vec{w}\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{N} e_i^2$$

$$\text{s.t.} \quad y_i[<\vec{w}, \phi(\vec{x}_i)> + b] = 1 - e_i, \quad i = 1,2,\ldots N \tag{1}$$

In LS-SVMs, the optimization problem can be solved by solving the following linear equation:

$$\begin{bmatrix} 0 & -\vec{y}^T \\ \vec{y} & \Omega + \gamma^{-1} I_N \end{bmatrix} \begin{bmatrix} b \\ \vec{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix} \tag{2}$$

where $\vec{y} = [y_1, \ldots y_N]^T$, $1_N = [1, \ldots 1]^T$, $\vec{\alpha} = [\alpha_1, \ldots \alpha_N]^T$; I_N is an $N \times N$ identity matrix; $\Omega \in R^{N \times N}$ is the kernel matrix defined by $\Omega_{ij} = y_i y_j \phi(\vec{x}_i)^T \phi(\vec{x}_j) = y_i y_j K(\vec{x}_i, \vec{x}_j)$; $N$is the sample number; $\vec{x}_i$ is the sample and y_i is its corresponding label; e_i is the error;