Methods 83 (2015) 3–17

Contents lists available at ScienceDirect

Methods

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

A novel mixed integer programming for multi-biomarker panel identification by distinguishing malignant from benign colorectal tumors

Meng Zou [a,1], Peng-Jun Zhang [b,1], Xin-Yu Wen [b], Luonan Chen [c,d,*], Ya-Ping Tian [b,*], Yong Wang [a,*]

[a] National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China
[b] Department of Clinical Biochemistry, State Key Laboratory of Kidney Disease, Chinese PLA General Hospital, Beijing 100853, China
[c] Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China
[d] School of Life Science and Technology, Shanghai Tech University, Shanghai 201210, China

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form 7 May 2015
Accepted 8 May 2015
Available online 15 May 2015

Keywords:
Translational bioinformatics
Multi-biomarker
Colorectal cancer
Mixed integer programming

ABSTRACT

Multi-biomarker panels can capture the nonlinear synergy among biomarkers and they are important to aid in the early diagnosis and ultimately battle complex diseases. However, identification of these multi-biomarker panels from case and control data is challenging. For example, the exhaustive search method is computationally infeasible when the data dimension is high. Here, we propose a novel method, MILP_k, to identify serum-based multi-biomarker panel to distinguish colorectal cancers (CRC) from benign colorectal tumors. Specifically, the multi-biomarker panel detection problem is modeled by a mixed integer programming to maximize the classification accuracy. Then we measured the serum profiling data for 101 CRC patients and 95 benign patients. The 61 biomarkers were analyzed individually and further their combinations by our method.

We discovered 4 biomarkers as the optimal small multi-biomarker panel, including known CRC biomarkers CEA and IL-10 as well as novel biomarkers IMA and NSE. This multi-biomarker panel obtains leave-one-out cross-validation (LOOCV) accuracy to 0.7857 by nearest centroid classifier. An independent test of this panel by support vector machine (SVM) with threefold cross validation gets an AUC 0.8438. This greatly improves the predictive accuracy by 20% over the single best biomarker. Further extension of this 4-biomarker panel to a larger 13-biomarker panel improves the LOOCV to 0.8673 with independent AUC 0.8437. Comparison with the exhaustive search method shows that our method dramatically reduces the searching time by 1000-fold. Experiments on the early cancer stage samples reveal two panel of biomarkers and show promising accuracy.

The proposed method allows us to select the subset of biomarkers with best accuracy to distinguish case and control samples given the number of selected biomarkers. Both receiver operating characteristic curve and precision-recall curve show our method's consistent performance gain in accuracy. Our method also shows its advantage in capturing synergy among selected biomarkers. The multi-biomarker panel far outperforms the simple combination of best single features. Close investigation of the multi-biomarker panel illustrates that our method possesses the ability to remove redundancy and reveals complementary biomarker combinations. In addition, our method is efficient and can select multi-biomarker panel with more than 5 biomarkers, for which the exhaustive methods fail.

In conclusion, we propose a promising model to improve the clinical data interpretability and to serve as a useful tool for other complex disease studies. Our small multi-biomarker panel, CEA, IL-10, IMA, and NSE, may provide insights on the disease status of colorectal diseases.

The implementation of our method in MATLAB is available via the website: http://doc.aporc.org/wiki/MILP_k.

* Corresponding authors at: Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China (L. Chen), Chinese PLA General Hospital, Beijing 100853, China (Y. Tian), and Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China (Y. Wang).

E-mail addresses: zoumeng@amss.ac.cn (M. Zou), zhangpj301@gmail.com (P.-J. Zhang), wendy36500@163.com (X.-Y. Wen), lnchen@sibs.ac.cn (L. Chen), tianyp61@gmail.com (Y.-P. Tian), ywang@amss.ac.cn (Y. Wang).

[1] Joint first authors.

http://dx.doi.org/10.1016/j.ymeth.2015.05.011
1046-2023/© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Colorectal cancer (CRC) is the second most common cause of cancer in women and the third most common cause in men. Every year over 1 million people get CRC resulting in 0.5 million deaths. It is the fourth most common cause of cancer death after lung, stomach, and liver cancers [1–3]. Five-year survival is 90% if the disease is diagnosed while still localized (i.e., confined to the wall of the bowel) but 70% for regional disease (i.e., disease with lymph node involvement) and only 12% if distant metastases are present [4,5]. Therefore, distinguishing CRC from benign colorectal diseases earlier is of great importance to improve the clinical success.

In recent years, many screening techniques have been developed to detect CRC to make CRC curable and preventable [6]. Unfortunately, screening compliance remains low, partly due to patients' dissatisfaction with fecal/endoscopic testing [7]. For example, screening by fecal occult blood test (FOBT) has been shown to significantly reduce CRC mortality but suffer from both false positive and false negative results [8–11]. Serum-based molecule biomarkers are highly attractive for early CRC screening as they could be easily integrated in any health checkup. Also it is not necessary to make special inspection if we can get accurate diagnosis only from serum, which would be significant for early diagnosis [12]. The European Group on Tumor Markers (EGTM) published guidelines on the use of tumor markers in CRC and focused almost exclusively on serum, especially carcinoembryonic antigen (CEA). However, none of these biomarkers has sufficient sensitivity and specificity for CRC [13–15]. Considering the specificity among samples and the complexity of colorectal disease, multi-biomarker panel is in pressing need. It combines several biomarkers together to achieve synergy effects and turns to be an appealing concept to distinguish CRC from benign colorectal diseases with a high accuracy.

The rationale behind multi-biomarker panel is that complex diseases, such as CRC, generally result from the intricate interactions among genetic, environmental, and lifestyle factors. So naturally at the microscopic level, CRC is typically caused by a combination of molecular perturbations and their interplay. It is also well-known that molecules, such as genes or proteins within a cell, do not function alone. They interact with each other to form networks or pathways for carrying out complex biological functions, so called network biomarker and dynamical network biomarker. Therefore, multi-biomarker panel has been intensively studied in recent years. It is a group of co-functional biomarkers and holds the great promise to early diagnosis of CRC [16–19].

However, identifying multi-biomarkers is not an easy task since the dimension of measured data is usually high and is increasing in genomics era. For example, given n biomarkers measured in serum, we aim to select p biomarkers as the optimal subset to achieve the best classification accuracy. Imagine that the exhaustive search method goes through all the possible combinations, for which the computation complexity would be $O(n^p)$. When n is large, say 10,000, the problem is computationally intensive. Thus, some heuristic methods have been proposed for such tasks. For example, minimum redundancy maximum relevance (mRMR) used a greedy strategy. In this scenario the optimality cannot be guaranteed and the heuristic strategy will be easily trapped into local minimum and miss some critical combinations [20–22].

To meet the grand challenge, we propose a novel optimization model to directly minimize the classification error (maximize the classification accuracy) given the number of biomarkers k in the optimal multi-biomarker panel. This mixed integer programming model allows us to go through all the optimal combinations by varying parameter k from 1 to n. Moreover, we can check their accuracy and compare the selected combinations. In particular, an optimal multi-biomarker panel can be selected by balancing the parameter k and the classification accuracy. This selected multi-biomarker panel can be independently validated by other classifiers such as SVM. In this paper, we apply our method to distinguish malignant from benign colorectal tumors to identify multi-biomarker panel from clinical data for CRC diagnosis.

2. Materials and methods

2.1. Overview of data generation and analysis workflow

The schematic illustration of our procedure for multi-biomarker identification is shown in Fig. 1. We collect two groups of samples and measure the clinical data. Then based on nearest centroid classifier, we construct an optimization model for feature selection. By solving the formulated optimization problem, we select an optimal multi-biomarker panel with high accuracy. This panel can be used to help disease diagnosis.

2.2. Sample collection and clinical data generation

The study was approved by the Ethics Committee of the Chinese PLA General Hospital. All patients provided informed written consent for the study sample collection, as well as permission for their use in research. Peripheral blood samples (10 mL each) were collected in tubes that contained separating gel and clot activator. After centrifuging at 3400 rpm for 7 min, the supernatant was transferred into new tubes and the serum was aliquoted and stored $-80\,^{\circ}\mathrm{C}$ until detection. No freeze thawing was allowed prior to detection. The whole blood samples for colorectal disease were collected before surgery. In total there are 196 samples consisting of 101 CRC samples and 95 benign colorectal disease samples with diagnosis of colitis and colorectal polyp. The CRC stages are according to Dukes stage (Table 1). In addition, we collected these samples without co-morbidities by excluding diabetes, heart disease, and autoimmune disease.

For each patient, 61 biomarkers are measured for its concentration. The statistics and descriptions for these biomarkers are listed in Fig. 2. All the biomarkers mentioned above in our laboratory have been approved by ISO 15189 to ensure stable and comparable results.

2.3. Data pre-processing

Clinical data is always noisy and with outliers. It may influence the reliability of data analysis even when the number of outliers is small. Therefore, we pre-processed the data in the following way. We firstly sorted the samples for each biomarker in descending order; then we replaced the first 0–2.5% and the last 97.5%–1 biomarker values with the value at 2.5% and 97.5% point respectively. Secondly we normalized the data to make the mean and the standard deviation of each biomarker zero and one. Lastly we filtered out biomarker IL-4 since its values are almost equal to zero in all the samples.

2.4. AUC for single biomarker analysis

It is important to quantitatively assess whether a biomarker has the ability to help doctors to diagnose disease. We used area under the receiver operating characteristic (ROC) curve (AUC) to evaluate a single biomarker's ability in distinguish case and control [66]. For each biomarker, we sorted all the samples in descending order; then we took a real number as cutoff and supposed the samples located in the left side of the cutoff as the predicted CRC and the