Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 135 (2015) 185–191

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest

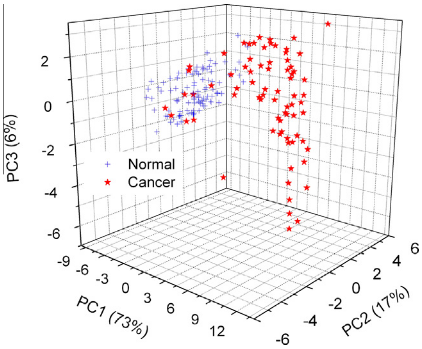
Hui Chen [a], Zan Lin [b], Hegang Wu [c], Li Wang [b], Tong Wu [b], Chao Tan [b,d,*]

[a] Hospital, Yibin University, Yibin, Sichuan 644007, China
[b] Department of Chemistry and Chemical Engineering and Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin, Sichuan 644007, China
[c] The First People's Hospital of Yibin , Yibin, Sichuan 644000, China
[d] Computational Physics Key Laboratory of Sichuan Province, Yibin University, Yibin, Sichuan 644007, China

HIGHLIGHTS

- Major spectral feature of colon tissues were captured.
- Random forest was used for constructing diagnostic models.
- Such a simulation procedure was fast and convenient.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form 16 June 2014
Accepted 2 July 2014
Available online 10 July 2014

Keywords:
Chemometrics
Spectrometry
Biodiagnostics
Random forest

ABSTRACT

Near-infrared (NIR) spectroscopy has such advantages as being noninvasive, fast, relatively inexpensive, and no risk of ionizing radiation. Differences in the NIR signals can reflect many physiological changes, which are in turn associated with such factors as vascularization, cellularity, oxygen consumption, or remodeling. NIR spectral differences between colorectal cancer and healthy tissues were investigated. A Fourier transform NIR spectroscopy instrument equipped with a fiber-optic probe was used to mimic in situ clinical measurements. A total of 186 spectra were collected and then underwent the preprocessing of standard normalize variate (SNV) for removing unwanted background variances. All the specimen and spots used for spectral collection were confirmed staining and examination by an experienced pathologist so as to ensure the representative of the pathology. Principal component analysis (PCA) was used to uncover the possible clustering. Several methods including random forest (RF), partial least squares-discriminant analysis (PLSDA), *K*-nearest neighbor and classification and regression tree (CART) were used to extract spectral features and to construct the diagnostic models. By comparison, it reveals that, even if no obvious difference of misclassified ratio (MCR) was observed between these models, RF is preferable since it is quicker, more convenient and insensitive to over-fitting. The results indicate that NIR spectroscopy coupled with RF model can serve as a potential tool for discriminating the colorectal cancer tissues from normal ones.

Introduction

Cancer is a disease characterized by uncontrollable growth and differentiation of cells and is one of the principal causes of death in

* Corresponding author at: Department of Chemistry and Chemical Engineering and Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin, Sichuan 644007, China. Tel./fax: +86 831 3551080.
E-mail address: chaotan1112@163.com (C. Tan).

http://dx.doi.org/10.1016/j.saa.2014.07.005
1386-1425/© 2014 Elsevier B.V. All rights reserved.

the world [1–3]. Colorectal cancer is a type of cancer due to uncontrolled cell growth in the colon, rectum, or appendix. Despite the progress in diagnostic techniques, unfortunately, the vast majority of colorectal cancers, more than 90%, have been either advanced or metastasized by the time they are diagnosed. Hence, there is an urgent need to develop accurate, fast, convenient, and inexpensive diagnostic method to detect the malignancy in the earlier stage for increasing the survival probability [4].

Even if host genes, bacterial virulence and environmental factors have been observed in oncogenic process, the underlying molecular mechanism is still poorly understood [5]. Nowadays, there exist several available diagnostic methods for colorectal cancer [6]. However, based on the conventional screening methods such as a white-light endoscope, it is difficult to probe early neoplasia or subtle lesions. FOBT coupled with subsequent colonoscopy is a popular choice for early detection of colorectal cancer. Biopsy followed by pathological assessment remains the gold standard, but it involves a complex procedure composed of fixation, dehydration, embedding, slicing and staining. In short, conventional diagnostic methods are time-consuming, labor-intensive, require experienced experts, and are strongly dependent on the experts' ability and subjective judgment [7].

In recent years, optical spectroscopic methods have been considerably investigated for cancer and precancer diagnosis and evaluation [8–10]. Among these methods, near-infrared (NIR) spectroscopy has shown huge potential since it can provide rich information of the molecular composition and structures of biological tissues. It has been used in some cancer researches including stomach [11], lung [12], breast [13], cervix [14], prostate [15], etc. Also, NIR-related methods are reagent-free, can rapidly detect changes of cells and tissues at the molecular level, particularly during carcinogenesis. It is known that biological tissues usually comprise DNA/RNA, proteins, carbohydrates, lipids and water as the main constituents; all of these can contribute meaningfully to the NIR absorption profile and therefore provide the informative basis for diagnosing cancer [16].

Even if the NIR is defined as encompassing the spectral range of 780–2500 nm, it is convenient to subdivide further this region into short and long NIR subregions. The short NIR region mainly reflects the signal of heme proteins and cytochromes, and provides rich information about tissue blood flow, as well as oxygen saturation and consumption. Long NIR region is associated to the combinations and overtones of hydrogen containing groups and thus captures valuable information on the chemical composition of tissues. Cancerous tissues are different from normal ones in composition and histology. Therefore, any alteration in the composition of the tissue can be reflected in NIR spectrum and used for diagnostic purposes. Several research groups have confirmed the advantages of NIR spectroscopy for malignant studies in both animal and human tissues. However, despite the merits, the NIR spectrum is an overlapped, broad and weak signal without distinct signature of individual components [17,18]. Under this situation, it is necessary to apply appropriate modeling methods to extract the subtle valuable information for clinical application. The modeling methods are of great importance and directly determine if the NIR-based applications can success. More recently, the so-called "ensemble" strategy has attracted more attention in various fields [19–21]. The main advantage of ensemble is that it can increase the accuracy and robustness of the predictor by a cooperation of many individual predictors. Random forest (RF) [22], a relatively new ensemble-based modeling technique, has attracted increasing interest of researchers. It combines Breiman's bagging idea and the random selection of input variables. RF holds many attractive features including a small number of tunable parameters, automatic calculation of generalization errors and variable importance, automatic handling of missing data, insensitive to over-fitting.

In the present work, the qualitative NIR spectral differences between colorectal cancer and healthy tissues in surgically resected specimens were investigated. A Fourier transform NIR spectroscopy instrument equipped with a fiber-optic probe was used to mimic in situ clinical measurements. A total of 186 spectra were collected and preprocessed by standard normalize variate (SNV) for removing unwanted background variances. All the specimens and spots used for spectral collection were confirmed by an experienced pathologist so as to ensure the representative of the pathology. Principal component analysis (PCA) was used to discover the possible clustering. Several methods including RF, partial least squares-discriminant analysis (PLSDA), *K*-nearest neighbor models and classification and regression tree (CART) were used to extract spectral features and to construct the diagnostic models. By comparison, it reveals that, even if no obvious difference of misclassified ratio (MCR) was observed between these models, RF is preferable since it is quicker, more convenient and insensitive to over-fitting. The results indicate that NIR spectroscopy coupled with RF model can serve as a potential tool for discriminating the colorectal cancer tissues from normal ones.

Theory and methods

Random forest

Random forest (RF) is one of potential algorithm for building classifiers and has been first introduced by Breiman [22]. RF exhibits some attractive features such as a small number of tunable parameters, automatic calculation of generalization errors, automatic handling of missing values, scale invariance and strong resistance to overfitting. It is actually an ensemble of unpruned decision trees by injecting randomness in both selecting samples for the training set and selecting variables for best splitting. At first, bagging (bootstrap aggregating) was introduced to decision trees for adding randomness of selecting training samples. So, dissimilar trees can be generated by re-sampling with replacement. The sensitivities of constructed trees to the constituent of the training set was reduced. Amit and Geman extended Breiman's concept by adding the randomness in selecting the best variables for splitting at each node [23]. Furthermore, Dietterich proposed the concept of random split selection that trees grow with a random subset of the best *K* variables at each node [24]. These attempts and the ideas were combined and evolved to the RF algorithm.

As an ensemble of models, RF uses majority voting for classification tasks and averaging for regression to make a final prediction. Also, in the frame of ensemble, there exist two necessary and sufficient conditions for ensuring the ensemble to be superior to its members. That is, the member models must be better than random guessing and be appropriately diverse. Models can be considered diverse only if their errors on unseen data are uncorrelated. The RF algorithm proceeds as follows:

(1) From the training set of n samples, draw a bootstrap sample, i.e., randomly sampling with replacement.
(2) For the bootstrap sample, produce a tree based on the following modification: choose the best split among a randomly selected subset rather than all variables at each node. The tree is grown to the maximum size without pruning.
(3) Repeat the above steps until ntree (an approximate large number) CART models are grown. In other words, each tree corresponds to a particular bootstrap sample and a total of ntree bootstrap samples will be drawn from the training dataset.
(4) Predict the class membership/labels of new samples by majority vote of the predictions from all ntree outputs.