Biochimica et Biophysica Acta 1860 (2016) 2682–2687

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study☆

Keiron O'Shea [a], Simon J.S. Cameron [a,b], Keir E. Lewis [c], Chuan Lu [d], Luis A.J. Mur [a,*]

[a] Institute of Biological, Environmental and Rural Studies, Aberystwyth University, Aberystwyth, Wales SY23 3DA, UK
[b] Division of Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, W6 8RP, UK
[c] Department of Respiratory Medicine, Prince Philip Hospital, Llanelli, Wales SA14 8LY, UK
[d] Department of Computer Science, Aberystwyth University, Aberystwyth, Wales SY23 3DA, UK

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form 9 June 2016
Accepted 11 July 2016
Available online 14 July 2016

Keywords:
Lung cancer
Small vs non-small cell lung cancer
Sputum
Metabolomics
Biomarkers
Artificial neural networks

ABSTRACT

Background: Lung cancer (LC) is one of the leading lethal cancers worldwide, with an estimated 18.4% of all cancer deaths being attributed to the disease. Despite developments in cancer diagnosis and treatment over the previous thirty years, LC has seen little to no improvement in the overall five year survival rate after initial diagnosis.
Methods: In this paper, we extended a recent study which profiled the metabolites in sputum from patients with lung cancer and age-matched volunteers smoking controls using flow infusion electrospray ion mass spectrometry. We selected key metabolites for distinguishing between different classes of lung cancer, and employed artificial neural networks and leave-one-out cross-validation to evaluate the predictive power of the identified biomarkers.
Results: The neural network model showed excellent performance in classification between lung cancer and control groups with the area under the receiver operating characteristic curve of 0.99. The sensitivity and specificity of for detecting cancer from controls were 96% and 94% respectively. Furthermore, we have identified six putative metabolites that were able to discriminate between sputum samples derived from patients suffering small cell lung cancer (SCLC) and non-small cell lung cancer. These metabolites achieved excellent cross validation performance with a sensitivity of 80% and specificity of 100% for predicting SCLC.
Conclusions: These results indicate that sputum metabolic profiling may have potential for screening of lung cancer and lung cancer recurrence, and may greatly improve effectiveness of clinical intervention. This article is part of a Special Issue entitled "System Genetics" Guest Editor: Dr. Yudong Cai and Dr. Tao Huang.

1. Introduction

The year 2008 saw an estimated 12.7 million new cases of cancer, and 7.6 million cancer-related deaths worldwide [1]. While the incidence and mortality rates of most cancers is decreasing in the developed world, they are rising in emerging economies such as China and India. Migrant studies have found that cancer rates in the descendent generation of migrants tends to shift toward the host country, suggesting that environmental risk factors such as smoking and weight are responsible for the global variance in cancer rates [2].

1.1. Lung cancer

Lung cancer is a major cause of death in the developed and developing worlds. It is the leading cause of cancer-related deaths in men, and second only to breast cancer in women. There was an estimated 1.6 million new cases of lung cancer and 1.4 million deaths in 2008. This accounts for 12.6% of all cancer incidence and a staggering 18.4% of all cancer-related deaths [2]. This can be attributed to its poor prognosis, with the five-year survival rate being a mere 15%. Despite recent advances in lung cancer treatment, survival rates are low when compared to other forms of cancer [3]. However, improvements in surgical techniques and chemotherapy over the past twenty years has resulted in one-year lung cancer survival rates drastically improving. Despite this, the overall five-year lung cancer survival rates have remained stagnant at 6% for small cell lung cancer and 18% for non-small cell lung cancer. Unfortunately the vast majority (85%) of lung cancer cases are diagnosed at advanced stages, heavily reducing the effectiveness of treatment [1].

This can be attributed to the difficulty of effectively diagnosing cancer of the lung at stage early enough to make a real impact. One of the main difficulties is that symptoms of the conditions are often identical to less serious conditions. This makes the pre-clinical diagnosis of lung cancer particularly problematic as the observed symptoms are often confused with other respiratory conditions. Prognostic factors may help diagnose patients who show symptoms of a disease, or have an increased chance of recurrence or progression to advanced disease which should support clinicians in the creation of appropriate treatment

☆ This article is part of a Special Issue entitled "System Genetics" Guest Editor: Dr. Yudong Cai and Dr. Tao Huang.

* Corresponding author.
E-mail address: lum@aber.ac.uk (L.A.J. Mur).

http://dx.doi.org/10.1016/j.bbagen.2016.07.007
0304-4165/

plans. The World Health Organisation (WHO) have set out ten key principles to be met by an effective screening procedure in order for it to be beneficial and cost effective [4]. Currently there are no lung cancer screening techniques of which meet all of the ten conditions laid out by the WHO.

1.2. Metabolomic insights into lung cancer

An emerging screening methodology to other traditional screening methods is the utilisation of molecular biomarkers in biofluids. The ease of analysis of biofluids using mass spectrometry (MS) or nuclear magnetic resonance (NMR) makes metabolomics a well-suited methodology for the non-invasive detection of biomarkers in lung cancer. Current focus of metabolomics in lung cancer has been on the exploitation of serum, urine and tumour biopsies. For example, the analysis of serum using liquid chromatography (LC-MS) and gas chromatography (GC–MS) approaches have suggested a potential use for biomarkers of lung cancer. A small-scale pilot study sampling lung cancer patients before and after surgical intervention, alongside patients without lung cancer has suggested ten candidate biomarkers for lung cancer, including sphingosine, oleic acid and serine [5].

Sputum has been suggested as a potential biofluid source of biomarkers in lung cancer [3], [6]. Recent work has used Fourier Transform Infra-Red (FTIR) spectroscopy as a non-invasive method to detect lung cancer in sputum samples. This work concluded that FTIR was able to sufficiently distinguish between lung cancer and control samples, and effectively act as a non-invasive, high-throughput and cost-effective method for screening sputum samples from high-risk patients. Furthermore, it further validated the use of sputum as an effective biofluid for lung cancer screening [7].

1.3. Artificial neural networks

Artificial neural networks (ANNs) are a class of sophisticated computational modelling structures that are inspired by biological neurological systems, regarding how they are able to learn and process highly non-linear information. The past three decades have seen ANNs being widely used for biomedical decision support systems [8–12].

In general, an artificial neural network is formed of interconnected processing units, commonly referred to as neurons. Each neuron applies an activation function over the weighted sum of the incoming stimuli (or inputs), and generate an output signal, which could be the input signal for other neurons. Many different neural network architectures exist, in this paper we will focus on the popular feed-forward artificial neural network, in particular the multi-layer perceptron (MLP) [13], that usually consists of multiple layers of neurons - the input layer, one or more hidden layers and the final output layer, as illustrated in Fig. 1.

The design of network architectures involves setting the number of hidden layers, the number of neurons within each layer, the connections between them, and the type of activation function to use. The connection weights in the network could be adjusted through a learning algorithm that minimises the amount of error in the outputs compared to the true ones. Generalisation, and to avoid overfitting the training data would be a central issue both in network design and training.

2. Case study

We have recently developed this approach by employing flow-infusion electrospray-mass spectrometry (FIE-MS) to evaluate the potential of spontaneous sputum as a source of non-invasive metabolomic biomarkers for LC status [14]. Spontaneous sputum was collected and processed from 34 patients suspected of having LC, and 33 healthy controls. Of the 34 patients, 23 were subsequently diagnosed with LC (LC+) at various stages of disease progression. The clinical characteristics of all samples taken are summarised in Table 1.

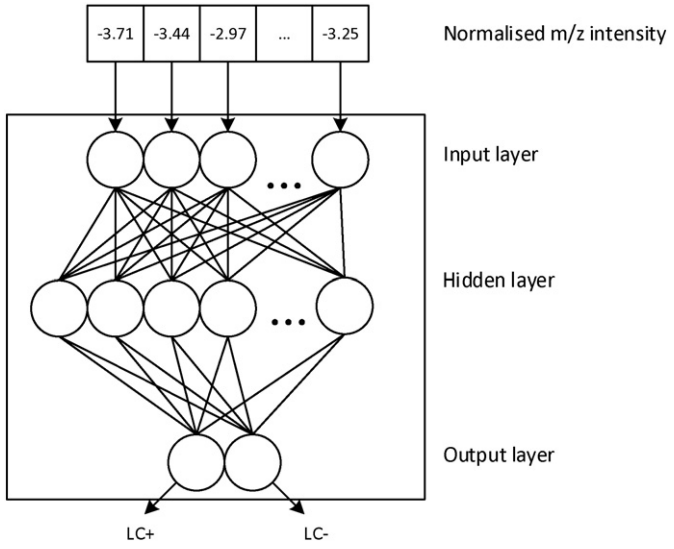

Fig. 1. Illustration of a three-layered feed-forward artificial neural network, where each neuron in one layer has connections to the subsequent layer.

In these preliminary analyses, discriminatory metabolites were identified using ANOVA and Random Forest and included Ganglioside GM1 which has previously been linked to lung cancer [15]. This suggested that the use of sputum as a non-invasive source of metabolite biomarkers may aid in the development of an at-risk population screening programme for lung cancer or enhanced clinical diagnostic pathways. We now demonstrate how further data-mining of the FIE-MS data has revealed further metabolite biomarkers, and evaluate further the use of metabolomics to yield biomarkers for distinguishing lung cancer type.

2.1. Ethics statement

The MedLung observational study (UKCRN ID 4682) received loco-regional ethical approval from the Hywel Dda Health Board (05/WMW01/75). Written informed consent was obtained from all participants at least 24 h before sampling, at a previous clinical appointment, and all data was link anonymised before analysis.

2.2. Mass spectrometry

Frozen sputum samples were thawed before being exposed to 0.5 mL of dithiothreitol (DTT) to isolate sputum cells. Each sample was mixed using a vortex mixer for 15 min before being centrifuged at 1800*g* for 10 min before removing the supernatant. Sputum pellets were then analysed using flow infusion electrospray ion mass spectrometry (FIE-MS).

Signals identified under 50 *m/z* were removed, and the resulting FIE-MS data matrix contains 2582 *m/z* values after binning. The data was further preprocessed by total ion count (TIC) normalisation (to ensure

Table 1
Summary of clinical characteristics of patients with lung cancer (LC+), symptoms (LC−) and healthy controls.

Characteristics	Lung cancer (LC+)	Symptoms (LC−)	Healthy controls
Number	23	11	33
Age (mean ± SD)	66.6 ± 8.1	66.5 ± 14.3	55.3 ± 14.6
Gender (male/female)	11/12	10/1	20/13
Smoking (current/ex/never)	10/10/3	3/0/8	15/18/0
Previous cancer (yes/no)	3/20	N/A	N/A
Final clinical diagnosis (SCLC/NSCLC/radiological)	5/17/1	N/A	N/A
CO level (ppm)	4.2 ± 2.8	3.7 ± 1.3	N/A