RESEARCH ARTICLE

Prediction of Lung Cancer Based on Serum Biomarkers by Gene Expression Programming Methods

Zhuang Yu[1&*], Xiao-Zheng Chen[1&], Lian-Hua Cui[2], Hong-Zong Si[3], Hai-Jiao Lu[1], Shi-Hai Liu[4]

Abstract

In diagnosis of lung cancer, rapid distinction between small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) tumors is very important. Serum markers, including lactate dehydrogenase (LDH), C-reactive protein (CRP), carcino-embryonic antigen (CEA), neurone specific enolase (NSE) and Cyfra21-1, are reported to reflect lung cancer characteristics. In this study classification of lung tumors was made based on biomarkers (measured in 120 NSCLC and 60 SCLC patients) by setting up optimal biomarker joint models with a powerful computerized tool - gene expression programming (GEP). GEP is a learning algorithm that combines the advantages of genetic programming (GP) and genetic algorithms (GA). It specifically focuses on relationships between variables in sets of data and then builds models to explain these relationships, and has been successfully used in formula finding and function mining. As a basis for defining a GEP environment for SCLC and NSCLC prediction, three explicit predictive models were constructed. CEA and NSE are requently-used lung cancer markers in clinical trials, CRP, LDH and Cyfra21-1 have significant meaning in lung cancer, basis on CEA and NSE we set up three GEP models-GEP 1(CEA, NSE, Cyfra21-1), GEP2 (CEA, NSE, LDH), GEP3 (CEA, NSE, CRP). The best classification result of GEP gained when CEA, NSE and Cyfra21-1 were combined: 128 of 135 subjects in the training set and 40 of 45 subjects in the test set were classified correctly, the accuracy rate is 94.8% in training set; on collection of samples for testing, the accuracy rate is 88.9%. With GEP2, the accuracy was significantly decreased by 1.5% and 6.6% in training set and test set, in GEP3 was 0.82% and 4.45% respectively. Serum Cyfra21-1 is a useful and sensitive serum biomarker in discriminating between NSCLC and SCLC. GEP modeling is a promising and excellent tool in diagnosis of lung cancer.

Keywords: Lung cancer - diagnosis - gene expression programming - biomarker - cyfra21-1

Asian Pac J Cancer Prev, **15 (21)**, 9367-9373

Introduction

Lung cancer is one of the most frequent types of cancer in the world, both in terms of incidence and mortality, leading to three million deaths annually and resulting in an enormous global health problem (Cho, 2007). Known as diseases caused by uncontrolled cell growth in one or two lungs, its primary symptoms include chest pain, shortness of breath, cough and coughing up blood. Lung cancer may spreads to other organs and spreading to the brain can cause headache, vomiting, psychosis, unilateral limb paresthesia and visual impairment. In spite of the progress and efforts we made in treatment and diagnosis of lung cancer patients, the diagnosis of overall survival at five years is only 15%, and more than 75% patients at the time of diagnosis is already presenting with advanced stage of disease when therapeutic options are very limited (Patz et al., 2007).

The histological types of lung cancer are complicated because there are many types of normal airway epithelial cells. In the process of tumor genesis, pluripotent stem cells can differentiate into different directions, so that there is a significant heterogeneity of lung cancer in the histological types. Based on histopathological presentation, lung cancer is sub-divided into four major histological subtypes: small cell lung cancer (SCLC), squamous cell carcinoma (SCC), adenocarcinoma (ADC), and large cell carcinoma (LCC). The latter three, collectively referred as non-small cell lung cancer (NSCLC). So, from clinical point of view, domestic and overseas, these four completely different types of lung cancer are divided into two categories: small cell lung cancer (SCLC) (16.8%) and non-small cell lung cancer (NSCLC) (80.4%) (Travis et al., 1995). Since the biological behavior of small cell lung cancer and other types of lung cancer are significantly different-namely, the former is clinically characterized by

*[1] Department of Oncology, [4] The central Laboratory, The Affiliated Hospital of Qingdao University, [2]Department of Public Health, Qingdao University Medical College, [3] Department of Pharmacy, Qingdao University, Institute for Computational Science and Engineering, Laboratory of New Fibrous Materials and Modern Textile, the Growing Base for State Key Laboratory, Qingdao, Shandong, China [&]Equal contributors *For correspondence: yuzhuang2002@163.com*

highly malignant, early widespread metastases, sensitivity to radiotherapy and chemotherapy, so the treatment of small cell lung cancer is different to other types of lung cancer. As different types of lung cancer have different treatments, the differential diagnosis in early stage is very important to improve the survival rate. Therefore, early diagnosis of lung cancer types may play a very important role in the improvement of therapeutic outcomes. Together with the advances in imaging studies (such as Computed Tomography, Whole-body positron-emission tomography, Chest Radiograph, Magnetic Resonance Imaging and Sputum Cytology) and endoscopic examinations (such as bronchoscope and mediastinoscope), the development of biomarkers useful for serum diagnosis is crucial for the early diagnosis of lung cancer.

Biomarkers are reflecting the chemical and biological substances presented in the tumor. They may not exist in normal adult tissue or found only in embryonic tissue. Their content in the tumor tissues could significantly exceed the content in normal organization. Their quantitative or qualitative change may indicate the nature of the tumor, so as to understand the tumor tissue, cell differentiation and cell function, to help the diagnosis, classification, prognosis judgment and treatment of tumor (Leidinger et al., 2010).

Detection of serum tumor marker levels becomes ways to improve the rate of early diagnosis of lung cancer (Zhang et al., 2013). Markers which have been widely used for diagnosing lung cancer include CEA (Yang et al., 2014), Cyfra21-1 (Tomita et al., 2010), NES (Chu et al., 2011), LDH (Ziaian et al., 2014), CRP (Onitilo et al., 2012) and other indexes. However, single marker of lung cancer shows low diagnostic specificity and sensitivity, having poor value for the diagnosis of lung cancer (Leidinger et al., 2010). Therefore, there is a consensus that a combination of markers can increase the diagnostic specificity and sensitivity than single marker (Schneider et al., 2002; Farlow et al., 2010; Wang et al., 2013; Li et al., 2014). For example, in diagnosing some malignant tumors, such as prostate or ovarian cancer, by combining multiple biomarkers data, the diagnostic performance was improved rapidly than individual serum biomarker (Amsellem-Ouazana et al., 2005; Kim et al., 2012). In recent years, people have done a lot of researches on statistical point of view, hoping to establish a kind of intelligent diagnosis model based on multiple tumor marker detection data, in order to overcome the influence of some subjective and individual factors, and improve the further tumor diagnosis.

Over the last decades, machine learning algorithms have shown to be a potential tool in medicine. The Naives Bayes algorithm and the support vector machine (SVM) (Gopinath and Shanthi, 2013), Fuzzy Logic and Artificial Neural Networks (ANN) (Biglarian et al., 2012) have been used as auxiliary tools in diagnosis and prognosis of cancer. These methods train as a classifier according to the features of essential biomarkers and other index. Then test the classifier in the same samples or the other samples. For example, Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer (Schneider et al., 2002), and ANN significantly improved the sensitivity of biomarkers in the Diagnosis of Lung Cancer (Feng et al., 2012).

As these machine learning methods are easily available as software packages and can achieve a highly accurate model of classification, they can be broadly adapted to predict lung cancer types by using input biomarker data. While, what is difficult for them is to explain how these data were used for classification. Moreover, little methods are able to select appropriate data automatically. So we can say the predictive results which without revealing the intrinsic link of each variable are "black box". Researchers tend to analyze the relationship between biomarkers and the lung cancer with statistical methods. So we proposed a new evolutionary algorithm -Gene expression programming (GEP).

Gene expression programming (GEP) is proposed by a Portuguese scientist called Candida Ferreira in 2000 and it is a new type of adaptive evolutionary algorithm based on biological structure and function. What it learns specifically is about the relationships between variables in different sets of data, and it builds models to describe these relationships. GEP does not require an accurate formulation of physical relationships, and can find the precise expression for different markers. The markers that show negative correlation with lung cancer will not be selected by the GEP classifier.

GEP is originated in the field of biology and developed from the genetic algorithm (GAs) and genetic programming (GP). It inherits the traditional advantages of GAs and GP, based on these advantages, a genetic operation was developed which is specific to GEP. A large number of experiments prove that the GEP algorithm and a variety of improved algorithms have a very good performance in using simple coding to solve complex problems.

On the other hand, GEP have played a significant role to predict essential proteins indispensable for cell survive (Zhong et al., 2013). GEP has been used to predict adverse events of radical hysterectomy in cervical cancer patients (Kusy et al., 2013). Moreover, a GEP was applied to automatically detect the population with fatty liver in computed the teat date of serum glucose, the total cholesterol, low density lipoprotein cholesterol, high density lipoprotein cholesterol and triglyceride from 196 power plant workers. However, there is no relevant literature researched that GEP is effective in auxiliary diagnosis of lung cancer.

In this study we record the level of an assortment of biomarkers which previously proved to have prognostic or diagnostic value of lung cancer, as a first step in the effort to improve the diagnostic accuracy of biomarkers and establish a novel multi-analyze serum biomarker test for prediction of lung cancer through the use of Gene Expression Programming .In this way can we develop a best artificial calculation model that can be widely used in lung cancer types predicting.

Materials and Methods

Subjects

We selected 180 patients who were diagnosed with

biopsy-proven untreated pulmonary malignant disease. They all hospitalized in the Affiliated Hospital of Qingdao University from January 2006 to September 2013. The histological subtypes in the 180 patients with lung cancer were NSCLC in 120, SCLC in 60 patients. Approval has been obtained from the relevant ethics committee and all participants were given the written informed consents. Histological diagnosis of primary lung cancer was established according to the revised classification of lung tumors of the World Health Organization and the International Association for Lung Cancer Study .A summary of the clinical characteristics of the subjects, together with a breakdown of each group by age, gender and more detailed information is presented in Table 1.

Biomarker selection

Our GEP classifier is constructed to predict lung cancer types based on various biomarkers. Because CEA and NSE are the most widely used biomarkers among these patients, we implement experiments based on data of both of them. C-reactive protein (CRP) is an acute-phase protein produced mainly by hepatocytes in the presence of inflammation and serum CRP concentrations were significantly higher in NSCLC patients compared to the SCLC controls (Lee et al., 2009). Serum LDH levels have been correlated with poor prognosis and resistance to chemotherapy and radiotherapy in various neoplastic diseases (Zhao et al., 2013). Serum Cyfra21-1 is one of the most important serum markers in the diagnosis of non-small cell lung cancer (NSCLC), especially squamous-cell carcinoma (Ono et al., 2013; Wang et al., 2013). So, based on CEA and NSE, we set up three GEP models -GEP1 (CEA, NSE, Cyfra21-1), GEP2 (CEA, NSE, LDH), GEP3 (CEA, NSE, CRP).

Table 1. Characteristics of the Study Subjects

Characteristics		NSCLC (n=120)	SCLC (n=60)
Age (years)	Mean± SD	57.62±10.92	59.70±7.90
	Range	21-80	41-77
Gender	Male	73	51
	Female	47	9
Smoking index	0-400	71	17
	≥400	49	43
Cough	Yes	89	48
	No	31	12
Expectoration	Yes	82	21
	No	38	39
Chest congestion	Yes	86	38
	No	34	22

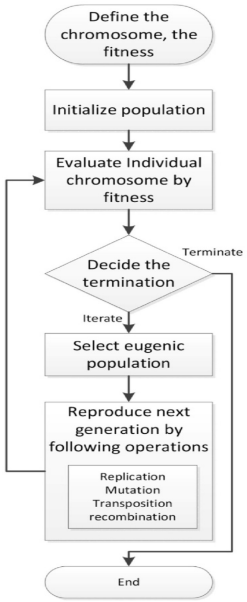

Figure 1. The Flowchart Showing the Entire Workflow of GEP Classifier

Measurement of serum biomarker concentrations

specimens: 10ml blood was collected from each patient in the fasting state, and processed immediately by centrifugation at 3000 rpm at room temperature for 10 min in a centrifuge.

Kits and Instruments: Levels of CEA, NSE and Cyfra21-1 were measured by using the Electrochemiluminecence immunoassay instrument. Their kits are from Roche Germany. Test them according to the standard operating procedure for kits. LDH and CRP are measured by using Olympus AU2700 automatic biochemical analyzer and auxiliary reagents. All operations are enforced according to the instruction strictly.

Normal Serum Reference Values: LDH: <245 U/L; CEA: <5.0 ng/mL; Cyfra21-1: <3.3 ng/mL; NSE: 15.7-17.1 ng/ mL; CRP: <8 ng/mL.

Statistical Analysis of individual serum biomarkers

Statistical analyses were performed using SPSS19.0. All data were expressed as mean±standard, groups comparison was conducted using analysis of variance (ANOVA). *P*-values<0.05 were considered significant difference, *P*>0.05 showed no significant difference among two groups.

Gene expression programming

Data preparation of GEP: Gene encodes the two types of lung tumors randomly. We code small cell lung cancer as 0, non-small cell lung cancer as 1. All data was entered into the computer twice. Original data are divided into 4 equal datasets, and one fold is used to train the classifier and the remaining three folds are used for testing the predictive performance of GEP. Since the ratio of SCLC and NSCLC in original data is about 1:3 (SCLC: NSCLC=60:180), each fold data maintains the same ratio of SCLC and NSCLC in original data. So there are 45 patients (SCLC: n=15 NSCLC: n=30) in testing set and 135 patients (SCLC: n=45 NSCLC: n=90) in training set. As we established three GEP models, the process is repeated three times to generated three classifiers, with each of the three datasets used exactly once as testing data.

GEP theory: To build a classifier of predicting lung cancer types using GEP, the following major steps are needed: defining a chromosome using a function and terminal set, initializing a population and generating a group of chromosomes, defining a fitness function for evaluating chromosomes, selecting eugenic ones from populations, reproducing a group of chromosomes of the next generation, and deciding the termination of the model. Figure 1 illustrates the flowchart of building GEP classifier.

In this work, GEP is developed to predict two essential

lung cancer types. First of all, a set of functions and terminals is chosen to define chromosomes that will be expressed as nonlinear entities. The set of functions contains arithmetic operators and logic operators, such as +, -, ×, ÷, min, max, equal, sqrt, log, exp, abs, while the set of terminals contains variables combination of serum biomarkers (for example: CEA+NSE+LDH and CEA+NSE+ Cyfra21-1) and relevant coefficients. Then we build the chromosomal structure. For each chromosome the length of its head and the number of genes will be given.

The second step is to randomly initialize population. The input parameter indicates the size of population. For example, when GEP is used to deal with CEA+NSE+Cyfra21-1, we vary the whole number of populations from 100 to 180, and keep track of results each model produces. According to the results, the performance of models gradually increases as the number of populations rises. Thus, we choose the maximum number, 180, as the input populations.

The third step is to define a fitness function to evaluate the individual chromosome. In order to obtain an optimal output, we select SSPN as our fitness function which is described as the product of sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). The formula is as follows: SSPNi=SNi × SPi × PPVi×NPVi. Where SN, SP, PPV, NPV are calculated respectively by using the following formulas for each chromosome: SNi=TPi/ (TPi+FPi), SPi=TNi/ (TNi+FNi), PPVi=TPi/ (TPi+FPi), NPVi=TNi/ (TNi+FNi)

The TNi, TPi, FNi and FPi are the numbers of true negatives, true positives, false negatives, and false positives, respectively. Given the content of 5 biomarkers, we use the fitness function to compute the scores of all chromosomes in population.

The forth step is to select the top 30% populations as the eugenic ones. Then the fifth step is that performing a set of genetic operations (including mutation, transposition and crossover) on eugenic ones reproduces chromosomes of the next generation that has the same size as former one.

Finally, we choose the maximum as the number of generations to decide the termination of the model. In this study, the parameters used in our GEP classifier are listed in Table 2. We develop the program that predicts lung cancer types based on GEP in C++ Language.

Results

Test result of individual serum tumor markers

SCLC and NSCLC groups are significantly different regarding to CEA, NSE, CPR, LDH and Cyfra21-1 concentrations in serum from 60 SCLC patients and 180 NSCLC subjects. The results of the measurements are as shown in Table 3.

Classification and prediction

Investigation of GEP models and their performance in diagnosing: We set up three GEP models. Observe their performances in diagnosing lung cancer types, then choose the one which obtains the highest prediction accuracy rate as the optimum GEP model.

The results of running the three GEP models were given in Table 4. It can be observed that in GEP3, 128 of 135 subjects in the training set as well as 41 of 60 subjects in the test set were classified correctly, its training set accuracy was 94.82% and the test set accuracy was 91.11%. By contrast, with the same function set, GEP1 represented the accuracy 94.07% and 77.78% while GEP2 represented 93.33% and 80.00% for training and testing set respectively. Compared with GEP1, the accuracy was significantly decreased by 1.5% and 6.6% in training set and test set in GEP2, in GEP3 was 0.82% and 4.45% respectively. It was concluded that the most accurate model was GEP3. The combination of CEA, NSE and Cyfra21-1 has the optimum predicting performance.

The expressions of our GEP models: The biomarkers were assigned to the columns as independent input

Table 2. Parameters Used in Our GEP Method

Parameter	Description of Parameter	setting of parameter
P1	Function set	+ - * / Exp Log Logi
P2	Number of genes	5
P3	Head Size	8
P4	Linking Function	Addition
P5	Number of chromosomes	50
P6	Number of generation	200
P7	Number of genes	5
P8	Number of Tries	3
P9	Max. Complexity	5
P10	Mutation rate	0.044
P11	IS Transposition Rate	0.1
P12	RIS Transposition Rate	0.1
P13	Gene Transposition Rate	0.1
P14	One-Point Recombination Rate	0.3
P15	Two-Point Recombination Rate	0.3
P16	Gene Recombination Rate	0.1

Table 3. Comparison of 5 Biomarkers Distribution in SCLC and NSCLC

Biomarker	NSCLC(n=120) Mean ± Standard	SCLC(n=60) Mean ± Standard	*P*-value
LDH	161.290±62.179	209.880±161.322	>0.005
CRP	25.079±24.817	14.935±21.078	<0.001
CEA	51.493±77.529	25.074±40.957	<0.001
NSE	13.638±5.571	62.972±63.012	<0.001
CYFRA 21-1	12.447±15.814	6.418±9.567	<0.001

*Compare Small cell lung cancer group and non-small cell lung cancer group, the differences of CEA, NSE, CRP and Cyfra21-1.were statistically significant <0.01, the differences of LDH were not statistically significant: P>0.05.

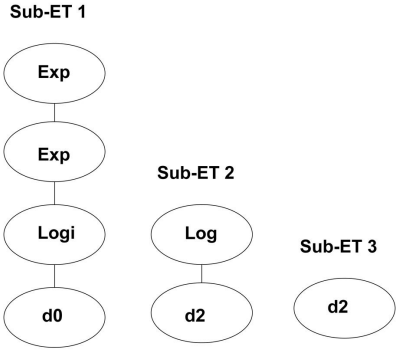

Figure 2. Expression Tree

Table 4. The Accuracy of GEP Models in Predicting Lung Cane Types

GEP model	Combination	Training set (n=135)		Testing set (n=45)	
		Best fitness	Accuracy	Best fitness	Accuracy
GEP1	1.CEA+NSE+LDH	127	94.07%	41	77.78%
GEP2	2.CEA+NSE+CRP	126	93.33%	36	80.00%
GEP3	3.CEA+NSE+ Cyfra21-1	128	94.82%	41	91.11%

Table 5. Statistics of Test Set and Training Set

Statistics	Training	Test
General Information		
Best Fitness:	128	41
Max. Fitness:	135	45
R-square:	0.78	0.65
Additional Information		
Calc. Errors:	0	0
Accuracy:	0.9482	0.9111
Error:	0.0518	0.0889
Sensivity:	1	1
Specificity:	0.8444	0.7333
PPV:	0.9278	0.88235
NPV:	1	1
Correlation Coefficient (CC):	0.88	0.88
Mean Squared Error (MSE):	0.05	0.05
Root Mean Squared Error (RMSE):	0.23	0.23
Relative Absolute Error (RAE):	0.12	0.12
Mean Absolute Error (MAE):	0.05	0.052
Relative Squared Error (RSE):	0.23	0.23
Root Relative Squared Error (RRSE)	0.4	0.4

variables and "+ - */Exp Log Logi" were used as function set. Therefore, a mathematical model of output variables was developed by using GEP. The program was coded in C++. The obtained formulations correspond to the following equations are: i) GEP1 (CEA+NSE+LDH) model is: y=apsLogid[1]+apsLogi{[d[0]÷log10 (d[1])]+d[0]-d[2])}+apsLogid{[log10 (d[0])÷$e^{d[1]}$]+apsLogi[log10 (d[1])]}-[d[0]×d[1]+log10 (d[0])]÷$d[1]^2$+[d[0]×d[2]+$d[0]^2$]÷$d[2]^2$-d[2]. Where, d[0], d[1], d[2] were LDH, CEA, NSE respectively; ii) GEP2 (CEA+NSE+CRP) model is: y=log10{ (d[0]) (d[1])+ ($e^{e^{apsLogid[0]}}$)}-{[apsLogi ($e^{d[0]}$)]+[apsLogid[0]+d[1]]}+{d[1]×log10 ($e^{\log10\,(apsLogid[0])}$)-d[2]}. In this model, d[0], d[1], d[2] were CRP, CEA, NSE.

The expression of GEP classifier with the best prediction performance, GEP2 (CEA+NSE+CRP) model was obtained as: y=$e^{e^{apsLogid[0]}}$+{log10[d[0]×apsLogid (log10 (d[2]))]}-{d[1]+2d[2]+[log10 (d[2])]}. d[0], d[1], d[2] are CEA, NSE, Cyfra21-1 respectively.

Analysis of the best GEP model

Expression tree: In order to achieve the geneticization of function expression, we need to use the K-expression (from Kava language) to represent the gene fragment. The process of expression is shown in Figure 2. Table V shoes the statistics of test set and training set.

Discussion

Because of its characteristics such as high degree of malignancy and difficulty to detect it early, lung cancer leads to unsatisfactory outcomes and 5-year survival rate. Roughly divided into two groups according to pathology: NSCLC (80.4%) and SCLC (16.8%), lung cancer becomes one of the major diseases which seriously threat human's life and health (Travis et al., 1995; Welch et al., 2000). Patients with NSCLC are treated differently from those with SCLC. If the NSCLC can be diagnosed in early stage, it is possible that the patients would undergo surgery and may achieve healing, thereby reducing mortality. Therefore, the distinction between them is extremely important (Travis, 2011).

Many studies have considered to lung cancer types (Li et al., 2005; Taguchi et al., 2007; Kligerman and Abbott, 2010; Wrona and Jassem, 2010; Nevins, 2011; Raj et al., 2011; Travis, 2011; Barash et al., 2012). For example, CT scanning, sputum cytology, LDCT, routine chest radiograph and, most recently, molecular biomarkers have evaluated various approaches for lung cancer classification. Based on chest X-Ray or sputum cytology, early screening tests of high risk individuals have not shown improvement in disease-specific survival. Chest LDCT scan, which recent NCCN guidelines endorsed as a screening tool for lung cancer, has been proven as an effective tool for the early detection of resectable disease. However, questions remain regarding the definition of lung cancer patients: how long and how frequent to screen; and high cost and potential toxicity from radiation exposure. The abnormity of the peripheral blood tumor markers is often earlier than radiographic abnormality; therefore for patients have no clinical symptoms, any abnormal markers could have prompted significant.

The expression of biomarkers in lung cancers is useful in the diagnosis and clinical management of patients with lung cancer. Biomarkers provide insight into histogenesis, interrelationships, and biological behavior of lung tumors. People have found a lot of tumor markers which were valuable for lung cancer diagnosis, efficacy detection and relapse diagnosis, such as AFP, PSA and CA125, have been proven respectively to be effective in the screening of liver, prostate and ovarian cancers (Perkins et al., 2003). Due to the complexity of tissue origin of lung cancer, involving the cancerous process of multiple genes and the heterogeneity of tumor antigens expression, a single tumor biomarker can not reflect lung cancer biological characteristics very well (Wang et al., 2013). Therefore, the sensitivity of a single tumor marker for diagnosis of lung cancer is relatively low (Tureci et al., 2006). Combining tumor marker examination can improve the lung cancer diagnostic rate, namely, combinatory analysis of tumor markers is a new development direction of the experiment teaching as well as an active research topic recently. Regarding the importance and the need for finding a new simple effective method for lung cancer type detection, we proposed a GEP method to determine the biomarker combination that best discriminates NSCLC from SCLC subjects based on a variety of biomarkers obtained from the clinic and they are easily obtained in the impoverished region.

This chapter presents data on lung cancer detection, involving some of the most studied and interesting lung cancer biomarkers to date-Cyfra21-1, NSE, CEA, as well as markers in clinical application such as CRP, LDH. GEP1 trained with 3 biomarkers (CEA, NSE, Cyfra21-1), correctly classified 128 of 35 subjects in the training set as well as 40 of 45 subjects in the test set , acquires the accuracy of 94.82% and 88.89% in the training set and test set. The accuracy of classification rate was slightly reduced in GEP2 (CEA, NSE, LDH) and GEP3 (CEA, NSE, CRP) since one biomarker was changed by 1.5% and 6.6%, 0.82% and 4.45% in the training set and test set, respectively. It is acknowledged that accuracy in GEP1 is higher than GEP2 and GEP3, GEP1 gained the best performance which trained with Cyfra21-1, NSE, CEA. Cyfra21-1 which showed the most powerful result compared to the other tumor markers, revealed a distinctive potential to differentiate patients with NSCLC from SCLC.

Cyfra21-1, which is a polypeptide tumor marker, is produced by almost all human cells and designated circulating cytokeratin-19 fragment (Wieskopf et al., 1995; Molina et al., 2003; Nakata et al., 2004). It is a unique epitope from a polypeptide and abundantly elaborated following cell death, whose diagnostic utility and prognostic relevance have been demonstrated in stomach cancer, colorectal cancer, breast cancer, and cervical cancer (Gaarenstroom et al., 1995; Lee, 2013; Wang et al., 2013; Gwak et al., 2014). Having been classified into 20 subtypes based on isoelectric point and differences in the molecular mass which was determined by 2-dimensional electrophoresis, Cyfra21-1 is a useful auxiliary biomarker in the diagnosis of NSCLC, and it has been reported that it has particularly high specificity for the diagnosis of lung squamous cell carcinoma (Ono et al., 2013). Therefore based on NSE, CEA and Cyfra21-1, the GEP1 accuracy is outstanding than GEP2 and GEP3. Moreover, the clinical information such as age, gender, smoking, nodules, hemoptysis and other index, may improve the GEP performance.

In summary, GEP is a nonlinear method of artificial intelligence with good reproducibility, which makes full use of the thoughts of biological replication expression. Based on GEP, we find a superior biomarker combination to distinguish lung cancer types. The auxiliary mode is convenient, economy and can be widely used in poverty-stricken areas. However, with the emergence of new predictive tumor markers, we are confined no longer to a single or several determinate tumor markers joint detection, but to large sample size, high amount of information and large scale on gene and protein levels.

Acknowledgements

This work was supported by the department of Science & Technology of Shandong province (Contract No. 2012YD18042 and 2010GWZ20260).

References

Amsellem-Ouazana D, Younes P, Conquy S, et al (2005). Negative prostatic biopsies in patients with a high risk of prostate cancer. Is the combination of endorectal MRI and magnetic resonance spectroscopy imaging (MRSI) a useful tool? A preliminary study. *Eur Urol*, **47**, 582-6.

Barash O, Peled N, Tisch U, et al (2012). Classification of lung cancer histology by gold nanoparticle sensors. *Nanomedicine*, **8**, 580-9.

Biglarian A, Bakhshi E, Gohari MR, et al (2012). Artificial neural network for prediction of distant metastasis in colorectal cancer. *Asian Pac J Cancer Prev*, **13**, 927-30.

Cho WC (2007). Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. *Biomed Pharmacother*, **61**, 515-9.

Chu XY, Hou XB, Song WA, et al (2011). Diagnostic values of SCC, CEA, Cyfra21-1 and NSE for lung cancer in patients with suspicious pulmonary masses: a single center analysis. *Cancer Biol Ther*, **11**, 995-1000.

Farlow EC, Vercillo MS, Coon JS, et al (2010). A multi-analyte serum test for the detection of non-small cell lung cancer. *Br J Cancer*, **103**, 1221-8.

Feng F, Wu Y, Wu Y, et al (2012). The effect of artificial neural network model combined with six tumor markers in auxiliary diagnosis of lung cancer. *J Med Syst*, **36**, 2973-80.

Gaarenstroom KN, Bonfrer JM, Kenter GG, et al (1995). Clinical value of pretreatment serum Cyfra 21-1, tissue polypeptide antigen, and squamous cell carcinoma antigen levels in patients with cervical cancer. *Cancer*, **76**, 807-13.

Gopinath B, Shanthi N (2013). Support Vector Machine based diagnostic system for thyroid cancer using statistical texture features. *Asian Pac J Cancer Prev*, **14**, 97-102.

Gwak HK, Lee JH, Park SG (2014). Preliminary evaluation of clinical utility of CYFRA 21-1, CA 72-4, NSE, CA19-9 and CEA in stomach cancer. *Asian Pac J Cancer Prev*, **15**, 4933-8.

Kim YW, Bae SM, Lim H, et al (2012). Development of multiplexed bead-based immunoassays for the detection of early stage ovarian cancer using a combination of serum biomarkers. *PLoS One*, **7**, 44960.

Kligerman S, Abbott G (2010). A radiologic review of the new TNM classification for lung cancer. *AJR Am J Roentgenol*, **194**, 562-73.

Kusy M, Obrzut B, Kluska J (2013). Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients. *Med Biol Eng Comput*, **51**, 1357-65.

Lee JG, Cho BC, Bae MK, et al (2009). Preoperative C-reactive protein levels are associated with tumor size and lymphovascular invasion in resected non-small cell lung cancer. *Lung Cancer*, **63**, 106-10.

Lee JH (2013). Clinical Usefulness of Serum CYFRA 21-1 in Patients with Colorectal Cancer. *Nucl Med Mol Imaging*, **47**, 181-7.

Leidinger P, Keller A, Heisel S, et al (2010). Identification of lung cancer with high sensitivity and specificity by blood testing. *Respir Res*, **11**, 18.

Li J, Chen P, Mao CM, et al (2014). Evaluation of diagnostic value of four tumor markers in bronchoalveolar lavage fluid of peripheral lung cancer. *Asia Pac J Clin Oncol*, **10**, 141-8.

Li J, Orlandi R, White CN, et al (2005). Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin Chem*, **51**, 2229-35.

Molina R, Filella X, Auge JM, et al (2003). Tumor markers (CEA, CA 125, CYFRA 21-1, SCC and NSE) in patients with non-small cell lung cancer as an aid in histological diagnosis and prognosis. Comparison with the main clinical and pathological prognostic factors. *Tumour Biol*, **24**, 209-18.

Nakata B, Takashima T, Ogawa Y, et al (2004). Serum CYFRA

21-1 (cytokeratin-19 fragments) is a useful tumour marker for detecting disease relapse and assessing treatment efficacy in breast cancer. *Br J Cancer*, **91**, 873-8.

Nevins JR (2011). Pathway-based classification of lung cancer: a strategy to guide therapeutic selection. *Proc Am Thorac Soc*, **8**, 180-2.

Onitilo AA, Engel JM, Stankowski RV, et al (2012). High-sensitivity C-reactive protein (hs-CRP) as a biomarker for trastuzumab-induced cardiotoxicity in HER2-positive early-stage breast cancer: a pilot study. *Breast Cancer Res Treat*, **134**, 291-8.

Ono A, Takahashi T, Mori K, et al (2013). Prognostic impact of serum CYFRA 21-1 in patients with advanced lung adenocarcinoma: a retrospective study. *BMC Cancer*, **13**, 354.

Patz EF, Jr., Campa MJ, Gottlin EB, et al (2007). Panel of serum biomarkers for the diagnosis of lung cancer. *J Clin Oncol*, **25**, 5578-83.

Perkins GL, Slater ED, Sanders GK, et al (2003). Serum tumor markers. *Am Fam Physician*, **68**, 1075-82.

Raj V, Bajaj A, Entwisle JJ (2011). Implications of new (seventh) TNM classification of lung cancer on general radiologists-a pictorial review. *Curr Probl Diagn Radiol*, **40**, 85-93.

Schneider J, Bitterlich N, Velcovsky HG, et al (2002). Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. *Int J Clin Oncol*, **7**, 145-51.

Taguchi F, Solomon B, Gregorc V, et al (2007). Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J Natl Cancer Inst*, **99**, 838-46.

Tomita M, Shimizu T, Ayabe T, et al (2010). Prognostic significance of tumour marker index based on preoperative CEA and CYFRA 21-1 in non-small cell lung cancer. *Anticancer Res*, **30**, 3099-102.

Travis WD (2011). Classification of lung cancer. *Semin Roentgenol*, **46**, 178-86.

Travis WD, Travis LB, Devesa SS (1995). Lung cancer. *Cancer*, **75**, 191-202.

Tureci O, Mack U, Luxemburger U, et al (2006). Humoral immune responses of lung cancer patients against tumor antigen NY-ESO-1. *Cancer Lett*, **236**, 64-71.

Wang WJ, Tao Z, Gu W, et al (2013). Clinical observations on the association between diagnosis of lung cancer and serum tumor markers in combination. *Asian Pac J Cancer Prev*, **14**, 4369-71.

Welch HG, Schwartz LM, Woloshin S (2000). Are increasing 5-year survival rates evidence of success against cancer? *JAMA*, **283**, 2975-8.

Wieskopf B, Demangeat C, Purohit A, et al (1995). Cyfra 21-1 as a biologic marker of non-small cell lung cancer. Evaluation of sensitivity, specificity, and prognostic role. *Chest*, **108**, 163-9.

Wrona A, Jassem J (2010). [The new TNM classification in lung cancer]. *Pneumonol Alergol Pol*, **78**, 407-17.

Yang ZM, Ding XP, Pen L, et al (2014). Analysis of CEA expression and EGFR mutation status in non-small cell lung cancers. *Asian Pac J Cancer Prev*, **15**, 3451-5.

Zhang D, Ren WH, Gao Y, et al (2013). Clinical significance and prognostic value of pentraxin-3 as serologic biomarker for lung cancer. *Asian Pac J Cancer Prev*, **14**, 4215-21.

Zhao D, Xiong Y, Lei QY, et al (2013). LDH-A acetylation: implication in cancer. *Oncotarget*, **4**, 802-3.

Zhong J, Wang J, Peng W, et al (2013). Prediction of essential proteins based on gene expression programming. *BMC Genomics*, **14**, 7.

Ziaian B, Saberi A, Ghayyoumi MA, et al (2014). Association of high LDH and low glucose levels in pleural space with HER2 expression in non-small cell lung cancer. *Asian Pac J Cancer Prev*, **15**, 1617-20.