Genetics and Molecular Research

Analysis of specific serum markers of colon carcinoma using a Bhattacharyya-based support vector machine

W.Y. Yang, G. Shi, L.P. Wu, S.T. Wei, Y.N. Huang, L.X. Tan, R.Z. Yang, C.X. Yan, E.T. Guo, H.Y. Wang, J.Z. Tong, Y. Dong and D.Z. Han

Department of Gastroenterology,
The First Affiliated Hospital of Henan University, Kaifeng, Henan Province, China

Corresponding author: D.Z. Han
E-mail: lijiansheng0256@163.com

Genet. Mol. Res. 16 (1): gmr16019521
Received November 7, 2016
Accepted November 7, 2016
Published January 23, 2017
DOI http://dx.doi.org/10.4238/gmr16019521

ABSTRACT. We aimed to evaluate the specificity of 12 tumor markers related to colon carcinoma and identify the most sensitive index. Bhattacharyya distance was used to evaluate the index. Then, different index combinations were used to establish a support vector machine (SVM) diagnosis model of malignant colon carcinoma. The accuracy of the model was checked. High accuracy was assumed to indicate the high specificity of the index. The Bhattacharyya distances of carcinoembryonic antigen, neuron-specific enolase, alpha-feto protein, and CA724 were the largest, and those of CYFRA21-I, CA125, and UGT1A83 were the second largest. The specificity of the combination of the above seven indexes was higher than that of other combinations, and the accuracy of the established SVM identification model was high. Using Bhattacharyya distance detection and establishing an SVM model

based on different serum marker combinations can increase diagnostic accuracy, providing a theoretical basis for application of mathematical models in cancer diagnosis.

Key words: Colon carcinoma; Tumor marker; Specificity; Bhattacharyya distance; Support vector machine

INTRODUCTION

Colon carcinoma is a common type of malignant tumor of the alimentary system. In recent years, as the daily diet of many individuals has changed, the incidence and death rate associated with colon carcinoma have increased worldwide. In America, the incidence of colon carcinoma has increased dramatically, making this cancer type the third highest among common malignant tumors (Levin et al., 2008). Because the onset of colon carcinoma is insidious with ambiguous symptoms, the opportunity for early treatment is often missed, and the patients are usually diagnosed when the cancer has progressed to middle or late stages (Onouchi et al., 2008). Therefore, the early diagnosis of colon carcinoma is particularly important for the management of this disease.

There are several diagnostic methods for detection of colorectal cancer. For example, fecal occult blood tests (FOBTs) can be used to identify fecal occult blood, one of the symptoms of early-stage colon carcinoma. FOBT is a common method used to detect colon carcinoma. Additionally, endoscopy performed using a fibrocolonoscope or electronic colonoscope, is commonly used in the clinical setting. This method can be used to directly observe colonic lesions and perform qualitative biopsy (Young and Cole, 2007). Endoscopy is the most effective means to identify and diagnose colon carcinoma and is also the most common and accurate method for early-stage diagnosis. Tumor marker tests are commonly used to diagnose all tumors. To date, carcinoembryonic antigen (CEA) is broadly used as a cancer marker in the clinical setting. Additionally, some carbohydrate antigens are evaluated as indexes of early-stage colon carcinoma; these antigens include CA199, CA242, and CA50. These indexes alone or their combinations are helpful for the early diagnosis of colon carcinoma in the clinical setting. Gene-based diagnosis is also used for detection of colon carcinoma. Colon carcinoma is a multigenic disease involving several carcinogenic steps. The occurrence and development of colon carcinoma involve changes in multiple cancer-associated genes. Mutations in genes such as *APC*, *KRAS*, *p53*, and *DCC* can occur during the process of carcinogenesis and metastasis (Oving and Clevers, 2002). Finally, enzymes such as telomerase (TLMA) (Hauguel and Bunz, 2003) and cyclooxygenase 2 can be used as markers of colon carcinoma.

Serum tumor markers usually refer to the substances in blood produced and released by tumor tissues. Analysis of tumor markers has been broadly applied in the clinical setting; however, this method has several limitations (Kawamura, 1996). The optimal serum tumor markers are sensitive and specific. However, there are many serum tumor markers used for detection of colon carcinoma, most of which are not sensitive or specific (Ochi et al., 1997). Thus, analysis of the sensitivity and specificity of tumor markers is clinically meaningful. In this report, we present a combined mathematical and bioinformatic analysis of the specificity of a few common tumor markers.

MATERIAL AND METHODS

General data

Group with Colon Carcinoma: A total of 100 patients who visited the Digestive Department of the First Hospital affiliated to the Henan University within the period from January 2013 to December 2013 and underwent surgery for colon carcinoma, including 56 men and 44 women (average age: 59.0 years, range: 25-82 years), were enrolled in this study. According to the World Health Organization standards on pathological types and degrees of differentiation, there were 72 cases of colorectal tubular adenocarcinoma, 17 cases of mucinous adenocarcinoma, and 11 cases of papillary-tubular adenocarcinoma. There were 12 cases exhibiting poorly differentiated tumors, 79 cases exhibiting moderately differentiated tumors, and nine cases exhibiting well-differentiated tumors. All patients were confirmed by operation and pathology, and imaging and surgical exploration demonstrated that no patients showed metastasis to other tissues or organs.

Control Group with Benign Tumors: Fifty patients who were admitted to our hospital within the same time period, including 21 men and 29 women (average age: 52.5 years, range: 32-80 years) were also enrolled in this study. There were 20 cases of colitis, 16 cases of polyposis coli, nine cases of colorectal tubular adenoma, and five cases of rectal-villous-papillary epithelioma. All diagnoses were proven by clinical analysis, endoscopy, and pathological examination.

All patients agreed to participate in the study and provided written informed consent.

Clinical examination of tumor markers

All patients were phlebotomized after fasting, and 2 mL isolated serum was cryopreserved at -80°C, being prepared for centralized serum examination. The levels of CEA, neuron-specific enolase (NSE), heat-shock protein 60 (HSP60), CYFRA21-I, tissue plasminogen activator (TPA), alpha-feto protein (AFP), CA199, CA242, CA724, CA125, CA153, and UGT1A8 in the serum were measured by enzyme-linked immunosorbent assays and a COBAS 6000 automatic electrochemiluminescence immunoassay analyzer (Roche, Switzerland).

Screening indexes using Bhattacharyya distances

Bhattacharyya distances were used to sequence and screen the indexes. Bhattacharyya distances show the upper bounds of the minimum error rate of Bayes in sample normal distributions. This method is linked to error rate, and it can theoretically gain the advantageous features of classifications but hardly obtain analytic solutions. For selection of features, multidimensional and low-dimensional data were both feasible. The definition of the Bhattacharyya distance of each index between colon carcinoma samples and normal samples is shown in formula 1 (Xuan et al., 2006). Larger Bhattacharyya distances were associated with better classified effects.

$$B_i = \frac{1}{4}\frac{(\mu_{i+} - \mu_{i-})^2}{({\sigma_{i+}}^2 + {\sigma_{i-}}^2)} + \frac{1}{2}\ln\left(\frac{{\sigma_{i+}}^2 + {\sigma_{i-}}^2}{2\sigma_{i+}\sigma_{i-}}\right) \quad \text{(Equation 1)}$$

In this formula, μ_{i+} and σ_{i+} are the mean and variance of colon carcinoma samples, respectively, and μ_{i-} and σ_{i-} are the mean and variance of the sample in the control group, respectively. In this study, the calculations for the Bhattacharyya distances were carried out using MATLAB.

Accuracy validation by support vector machine (SVM)

The specificity of indexes screened by Bhattacharyya distances was validated using SVMs, and the establishment, training, and validation of SVM models were all implemented based on the MATLAB tools program (Chang and Lin, 2011).

First, 150 patients were normalized. The malignant regions of samples were marked as 1, and the benign regions were marked as 0. Eighty of 100 patients with malignant tumors and 40 of 50 patients with benign tumors were chosen, yielding a matrix of 120 x 12. The samples were input into the SVM for training. During the training, penalty parameter C and nuclear parameter γ were gradually optimized to achieve better results. The remaining 20 patients with malignant tumors and 10 patients with benign tumors were evaluated as the testing samples and input into the SVM network after training; the corresponding results (1 or 0) were obtained. The accuracy could be determined by comparison with the objective.

RESULTS

Results of serum content analysis

The results from tumor marker analyses for the 150 samples in the two groups are listed in Table 1. The 12 indexes were CEA, NSE, HSP60, CYFRA21-I, TPA, AFP, CA199, CA242, CA724, CA125, CA153, and UGT1A8.

Table 1. Analysis of 12 serum markers in the two groups (means ± standard deviations).

Index groups	Colon cancer group	Control group
CEA	29.31 ± 831 (ng/mL)	4.28 ± 1.39 (ng/mL)
NSE	11.76 ± 2.33 (ng/mL)	2.45 ± 1.01 (ng/mL)
HSP60	587.29 ± 477.44 (pg/mL)	201.45 ± 120.97 (pg/mL)
CYFRA21-I	8.75 ± 2.22 (ng/mL)	1.98 ± 1.04 (ng/mL)
TPA	0.87 ± 1.25 (U/mL)	0.081 ± 0.54 (U/mL)
AFP	17.68 ± 5.15 (ng/mL)	2.78 ± 0.98 (ng/mL)
CA199	52.03 ± 38.34 (U/mL)	24.03 ± 12.22 (U/mL)
CA242	18.55 ± 10.09 (U/mL)	5.06 ± 1.47 (U/mL)
CA724	5.87 ± 1.25 (U/mL)	1.06 ± 0.77 (U/mL)
CA125	43.05 ± 9.73 (U/mL)	10.31 ± 7.65 (U/mL)
CA153	21.40 ± 8.63 (U/mL)	15.14 ± 2.83 (U/mL)
UGT1A8	8.52 ± 2.03 (ng/mL)	34.6 ± 12.16 (ng/mL)

Bhattacharyya distance of each index

The Bhattacharyya distance of each index was calculated according to formula 1. The results are shown in Table 2. The Bhattacharyya distances of NSE, CEA, CA724, and AFP were larger, followed by those of CYFRA21-I and CA125.

Table 2. Bhattacharyya distances of tumor markers between the two groups.

Index	CEA	NSE	HSP60	CYFRA21-I	TPA	AFP	CA199	CA242	CA724	CA125	CA153	UGT1A8
Bhattacharyya distance	3.4608	4.2107	1.2176	2.7314	0.9357	3.2135	1.0877	1.7578	3.4332	2.4567	1.0739	2.3742

Establishment of different diagnosis models by SVM

After optimization of the penalty parameter C and nuclear γ in the SVM model training, the optimal parameter (C, γ) was determined as (12, 1).

First, all indexes were used to established an SVM diagnosis model. The testing results are shown in Figure 1. The accuracy, sensitivity, and specificity of the model were 22/30 (73.33%), 15/20 (75%), and 7/10 (70%), respectively.

Then, four indexes (CEA, NSE, CA724, and AFP) whose Bhattacharyya distances were over 3 were combined to establish an SVM diagnosis model. The testing results are shown in Figure 2. The classification accuracy, sensitivity, and specificity of the model were 24/30 (80%), 16/20 (80%), and 8/10 (80%), respectively.

The seven indexes (CEA, NSE, CYFRA21-I, AFP, CA724, CA125, and UGT1A8) whose Bhattacharyya distances were more than 2 were used to establish another SVM diagnosis model. The testing results are shown in Figure 3. The accuracy, sensitivity, and specificity were 26/30 (86.67%), 17/20 (85%), and 9/10 (90%), respectively.

The three SVM diagnosis models based on 12 indexes, four indexes whose Bhattacharyya distances were over 3, and seven indexes whose Bhattacharyya distances were over 2, were established and compared, and their classification accuracies were 73.33, 80, and 86.67%, respectively. Among the three models, the third model, namely, the SVM diagnosis model based on seven indexes (CEA, CA50, CYFRA21-1, CA199, CA724, CA125, and UGT1A8) had the highest classification accuracy and was therefore selected in this study.

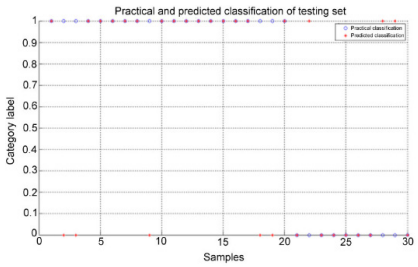

Figure 1. Analysis of the accuracy of the SVM model established using 12 tumor marker indexes.

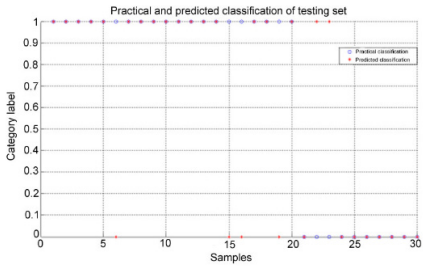

Figure 2. Analysis of the accuracy of the SVM model established using 5 tumor marker indexes.

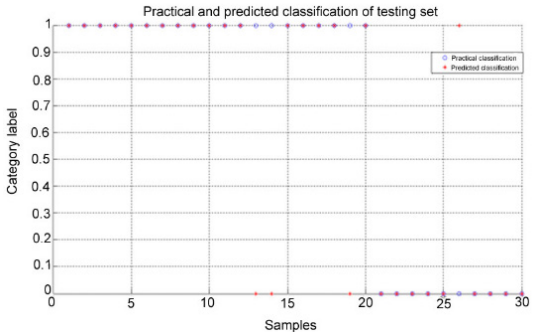

Figure 3. Analysis of the accuracy of the SVM model established using 7 tumor marker indexes.

DISCUSSION

There are many methods for analyzing the specificity of certain features. The most common method is the measurement of distance. Statistical pattern recognition states that as the distance between two categories becomes larger, the classification becomes easier, and the error rate becomes lower. Distance measures are also called class separability criteria or scatter criteria. The study of class separability in statistical pattern recognition is relatively deep. Distance is an important concept in statistical pattern recognition and is often analyzed using Euclidean distances, Mahalanobis distances, and Bhattacharyya distances (Li, 2009). Euclidean distances and Mahalanobis distances are defined in terms of space, whereas Bhattacharyya distances are defined in terms of probability. With regard to selection of features, subsets that can result in the largest classifying distance and the lowest error rate should be selected. The Bhattacharyya distance is usually applied to the feature analysis of gene expression profiles. Thus, Bhattacharyya distances are applicable to both multidimensional and low-dimensional data.

Tumor markers with high specificity screened out by Bhattacharyya distance included NSE, CEA, CA724, and AFP, which was consistent with a few previous studies. Several reports have demonstrated that CEA and CA724 are of high value in the diagnosis of colon cancer (Gebauer and Müller-Ruchholtz, 1997). A study by Wong (2006) found that the serum CEA contents in patients with colon cancer increased markedly, with a positive rate of 32.26%. Some studies have also reported that when CEA is used to diagnose alimentary canal neoplasms, colorectal cancer exhibits the highest positive rate (Kim et al., 2003). Chen et al. (2008) demonstrated that the sensitivity of AFP, NSE, CEA, and CA125 is 55.8% when these indexes are combined to detect gastric cancer and colon cancer. In addition, AFP is now recognized as the tumor marker with the highest specificity in primary liver cancer. However, 30-40% of samples are negative for APF (Jia et al., 2012). A study by Dai (2008) showed that AFP is statistically meaningless in the diagnosis of colon cancer. Notably, however, nonspecific serum tumor markers are of certain clinical value for diagnosis, assessment of lesion range and degree, evaluation of surgical outcomes, and examination of metastasis and postopeartive recurrence in patients with colon cancer (Yamamoto et al., 2001).

From the results of our study, we concluded that when 12 indexes were combined to establish an SVM model, the accuracy was 73.33%, which was not optimal. However, when seven indexes (CEA, NSE, CYFRA21-I, AFP, CA724, CA125, and UGT1A8) were combined to establish the SVM model, the accuracy was 86.67%. These data indicated that if

too many indexes were used, the effective indexes could be influenced by the redundant indexes, decreasing the accuracy. When four indexes (CEA, NSE, CA724, and AFP) whose Bhattacharyya distances were highest were combined to establish the SVM model, the accuracy was 80%, which was lower than that of the SVM model established by the seven indexes. A study by Fu et al. (2012) also found that when a single index was used to detect cancer, the difference between the indexes was not statistically significant. However, when five indexes, including CEA, CA199, CA724, and others were combined to detect cancer, the specificity and sensitivity were dramatically improved. Thus, these findings demonstrated that fewer indexes do not necessarily indicate better results but may lead to instability and unreliability of the results.

In summary, high test accuracy cannot be achieved using too many types of tumor marker indexes. The application of Bhattacharyya distances can effectively screen out indexes with high specificity, and the combination of specific indexes can be used to establish an SVM diagnosis model with high accuracy. However, it is not necessarily good to use only a few indexes. The number of indexes should be controlled properly to avoid occasionality of the results.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors thank all individuals who contributed to this study by providing advice and comments. Research supported by the National Natural Science Foundation of China (grant #81301963).

REFERENCES

Chang CC and Lin CJ (2011). LIBSVM: A library for support vector machines. *Acm. T. Intel. Syst. Tec* 2: 389-396.

Chen T, Su XX and Quan S (2008). Contrastive study on serum SGF, CEA, AFP, NSE, CA125 in clinical diagnosis of malignant tumors. The Seventh National Conference on laboratory medicine of Chinese Medical Association, Chongqing, 333.

Dai P (2008). The significance of serum tumor marker detection in colorectal cancer. Master's degree, Shanxi Medical University, Shanxi.

Fu HB, Wang WM and Cai QP (2012). The application of the combined test of tumor markers in colon carcinoma. *Chin. J. Clin. Ed.* 6: 5087-5090.

Gebauer G and Müller-Ruchholtz W (1997). Tumor marker concentrations in normal and malignant tissues of colorectal cancer patients and their prognostic relevance. *Anticancer Res.* 17 (4B): 2939-2942.

Hauguel T and Bunz F (2003). Haploinsufficiency of hTERT leads to telomere dysfunction and radiosensitivity in human cancer cells. *Cancer Biol. Ther.* 2: 679-684. http://dx.doi.org/10.4161/cbt.2.6.555

Jia BC, Luo XL, L R, Yue HF, et al. (2012). Diagnostic value of serum GP73 and AFP detection in primary hepatic carcinoma. *China J. Cancer Prev. Treat.* 19: 832-835.

Kawamura T (1996). [Current advancement of assay of tumor markers and the perspective in future]. *Nihon Rinsho* 54: 1642-1648.

Kim SB, Fernandes LC, Saad SS and Matos D (2003). Assessment of the value of preoperative serum levels of CA 242 and CEA in the staging and postoperative survival of colorectal adenocarcinoma patients. *Int. J. Biol. Markers* 18: 182-187.

Levin B, Lieberman DA, McFarland B, Andrews KS, et al. (2008). American Cancer Society Colorectal Cancer Advisory Group; US Multi-Society Task Force; American College of Radiology Colon Cancer Committee. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology* 134: 1570-1595. http://dx.doi.org/10.1053/j.gastro.2008.02.002

Li P (2009). Study on features gene selection of gastric cancer based on gene expression data. Master's thesis, Beijing University of Technology, Beijing.

Ochi Y, Okabe H, Inui T and Yamashiro K (1997). [Tumor marker--present and future]. *Rinsho Byori* 45: 875-883.

Onouchi S, Matsushita H, Moriya Y, Akasu T, et al. (2008). New method for colorectal cancer diagnosis based on SSCP analysis of DNA from exfoliated colonocytes in naturally evacuated feces. *Anticancer Res.* 28: 145-150.

Oving IM and Clevers HC (2002). Molecular causes of colon cancer. *Eur. J. Clin. Invest.* 32: 448-457. http://dx.doi.org/10.1046/j.1365-2362.2002.01004.x

Wong ZY (2006). The clinical significance of CEA, CA19-9 and CA242 detection of colorectal cancer, Master's thesis, Wenzhou Medical University.

Xuan GR, Zhu XM, Chai PQ, Zhang ZP, et a1. (2006). Feature Selection Based on the Bhattacharyya Distance. 18th International Conference on Pattern Recognition (ICPR'06), Hongkong, 4: 957.

Yamamoto H, Miyake Y, Noura S, Ogawa M, et al. (2001). [Tumor markers for colorectal cancer]. *Gan To Kagaku Ryoho* 28: 1299-1305.

Young GP and Cole S (2007). New stool screening tests for colorectal cancer. *Digestion* 76: 26-33. http://dx.doi.org/10.1159/000108391