

Support Vector Machine Approach for Cancer Detection using Amplified Fragment Length Polymorphism (AFLP) Screening Method

Waiming KONG, Lawrence THAM

Address: Bioinformatics Group, Nanyang Polytechnic,
180 Ang Mo Kio Ave 8
Singapore (589 830)

Email: KONG_WAI_MING@nyp.gov.sg, Lawrence_THAM@nyp.gov.sg

Abstract

Support Vector Machine is used to cluster data obtained from Amplified Fragment length Polymorphism screening of gastric cancer and normal tissue samples. Using the electrophoresis peak intensity measurements of the amplified fragments of the cancer and normal tissues, SVM was able to distinguish gastric cancer from normal tissue samples with a sensitivity of 0.98 and selectivity of 0.75. As AFLP is a low cost procedure which requires minimum prior sequence knowledge and biological material, SVM prediction of AFLP screening data is a potential tool for gastric cancer screening and diagnosis.

Keywords: SVM, AFLP, cancer detection

Introduction

Gastric carcinoma, despite its worldwide decline in incidence, is one of the leading causes of cancer mortality worldwide, second only to lung cancer among men and lung and breast cancer in women (Ferley et al, 2001). The high mortality rate of gastric cancer has been attributed mainly to late diagnosis of the disease and in part to the poor predictive value of current classification schemes in terms of clinical behavior of the disease. Surgery is the primary curative mode of treatment that has been shown to reduce mortality rate (Wu et al, 1997). Despite much progress in the use of chemotherapy and radiation in recent years, they remain palliative in nature (Crookes, 2002). In view of the high mortality and morbidity of gastric carcinoma, there's an urgent need to develop new diagnostics and treatment strategies for the disease.

Most screening efforts for gastric cancer, which relied on radiographic techniques such as photofluorography, are laborious, costly and clinician-dependent (Pisani et al, 1994; Fukao et al, 1995). Alternative methods that employ molecular techniques, including comparative genome hybridization (CGH), SNP and microsatellite genotyping, and expression microarrays, are currently being explored in the diagnosis and classification of gastric cancer. Although resolving power of these molecular assays have been shown to be more sensitive in the detection of biologically significant differences in tissues as

compared to more traditional approaches (Alizadeh et al, 2000; Bittner et al, 2000), they are not without drawbacks. CGH, with its limited mapping resolution (~2 Mbp), is only able to detect large chromosomal aberrations, while SNP and microsatellite screening methodologies require large amount of prior sequence knowledge. Expression microarrays, on the other hand, is marred by the amount of RNA required to generate an expression profile, as well as the inherent liability of RNA.

Amplified fragment length polymorphism (AFLP) screening is a DNA-based genotyping assay that has been predominantly used for strain typing in plants and bacteria (Vos *et al.*, 1995). It detects DNA restriction fragments by means of PCR amplification and required relatively little starting genetic material. In addition, AFLP is capable of surveying a target genome rapidly and comprehensively, without the need of prior sequence knowledge. Despite its wide use in the botanical and microbiological fields (Jonas *et al.*, 2000; Williams *et al.*, 2001), there are relatively few reports in which AFLP screening is applied on the human genome (Prochazka *et al.*, 2001).

In this paper, Support Vector Machine (SVM) is used to separate the AFLP data of cancer and the normal tissues. SVM, an effective method for general purpose supervised pattern recognition (Vapnik 1995, 1998), has been applied successfully to many biological data recently, including the identification of unknown genes using the gene expression data from DNA microarray hybridization experiments by Brown et al. (Brown et al., 2000), classification of ovarian cancer tissue (Furey et al., 2000) and prediction of protein structural test (Cai et al., 2001). In addition, SVM was also used for searching translation initiation sites (Zien et al., 2000) and for splice site recognition (Sonnenburg et al., 2002).

We investigated on the novel use of Amplified Fragment Length Polymorphism (AFLP) screening in the diagnosis and classification of cancers using a set of 58 gastric tumor and 16 normal genomic DNA samples. Here, SVM was used to cluster the AFLP data and the result shows that SVM can be used to differentiate cancer tissues from non-cancer tissues with a high level of accuracy. Jackknife test performed on the sample revealed sensitivity of 0.98 and specificity of 0.75.

Materials and Methods

Genomic DNA Sample

Genomic DNA samples of 58 gastric tumors were obtained from the Division of Medical Sciences, National Cancer Centre of Singapore. Normal genomic DNA samples were obtained from peripheral blood samples from healthy volunteers.

Amplified Fragment Length Polymorphism

AFLP is a DNA fingerprinting technique that detects DNA restriction fragments by means of selective PCR amplification. Genomic DNA is first digested by 2 restriction enzymes, a frequent-cutter endonuclease and a rare-cutter endonuclease. Double-stranded adapters are then ligated to the ends of the restriction fragments. Through two rounds of PCR amplification using DNA primers of increasing selectivity, the complexity of the genomic mixture is reduced, before separation of the final amplified fragments by electrophoresis. The unique banding patterns or ‘DNA fingerprint’ thus formed for a particular biological sample can be compared to ‘fingerprints’ of other samples. The AFLP screening performed was a modified version of the AFLP Analysis System I (Invitrogen) manufacturer’s instructions. Fragment analysis of AFLP products was performed by MegaBACE (Amersham Biosciences) capillary electrophoresis.

Data Generation

Electrophoretic peak data generated with Genetic Profiler (Amersham Biosciences) was normalized before clustering by SVM. Normalization of data across samples was performed using intensities of size standards (ET400-R standard) spiked into each well before electrophoresis. Normalization was used to correct for variations in sample intensity due to sample over- or under-loading due to small volume handling.

Accuracy of Diagnostic Tests

Accuracy of a diagnostic test can be expressed through sensitivity and specificity. Sensitivity refers to the ability of a certain diagnostic test to detect a particular disease. It is expressed as the probability of testing positive if the particular disease is truly present, i.e., the probability of having both a positive test and a positive diagnosis. Hence a test with 98% sensitivity means that 98% of those with the disease will test positive. Specificity, on the other hand, refers to the probability of testing negative if the disease is truly absent. In other words, 98% specificity means that 98% of those who are truly negative for the disease or problem will have a negative test while 2% of them will have a false positive test. See Table 1 for calculation.

Table 1. Calculation of Sensitivity and Specificity

		Disease	
		+	-
Test Results	+	TP	FP
	-	FN	TN

TP = number of true positive

FP = number of false positive

FN = number of false negative

TN = number of true negative

Sensitivity = $TP / (TP + FN)$

Specificity = $TN / (TN + FP)$

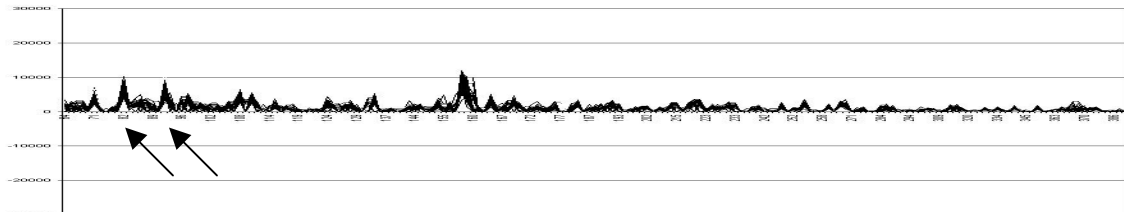
Support Vector Machines

SVM is a learning algorithm that was developed by Vapnik (Vapnik, 1995; Vapnik, 1998). The basic idea behind SVM is the formation of a linear classifier to separate the positive and negative training data. This linear classifier is then used on the test data. Depending on the position of the point falling on the positive or negative side of the linear classifier, the prediction result is calculated. In real world situation, a linear classifier may not be able to separate the positive and negative data sets. To overcome this limitation, kernels are used to transform the data into a feature space where the linear classifier can be used for the separation. Choosing the right kernel to use is therefore essential for the separation in the feature space.

Design and implementation

SVMLight (Joachims ,1999a; Joachims,1999b) was used to implement the SVM clustering. The jackknife test was used to examine the effectiveness of SVM to differentiate among the data. Jackknife test is also called the leave-one-out test, in which each data in the dataset is singled out as a tested data while all remaining data are used to train the SVM. One out of the 74 data was chosen to be the test data. The SVM was trained with the linear kernel on the remaining cancer and normal data as positive and negative data respectively. After training, the test data was used to determine the prediction of the SVM. This process was repeated for each of the 58 cancer data and 16 normal data. After the test of the entire data set, the sensitivity and the specificity of the result were calculated.

Cancer Sample



Normal Sample

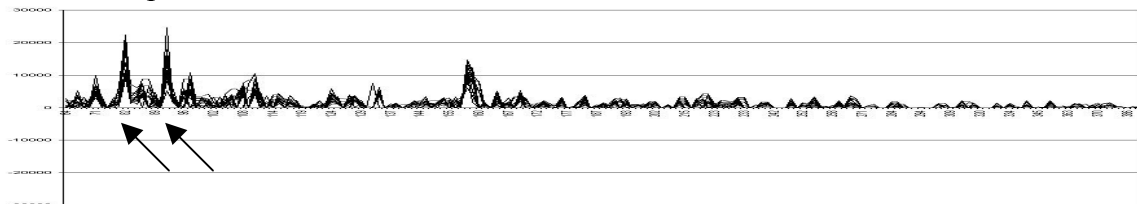


Fig. 1. Electrophoresis peak intensity versus molecular weight of AFLP data of cancer and normal samples with peaks at molecular weights 82 and 91 highlighted.

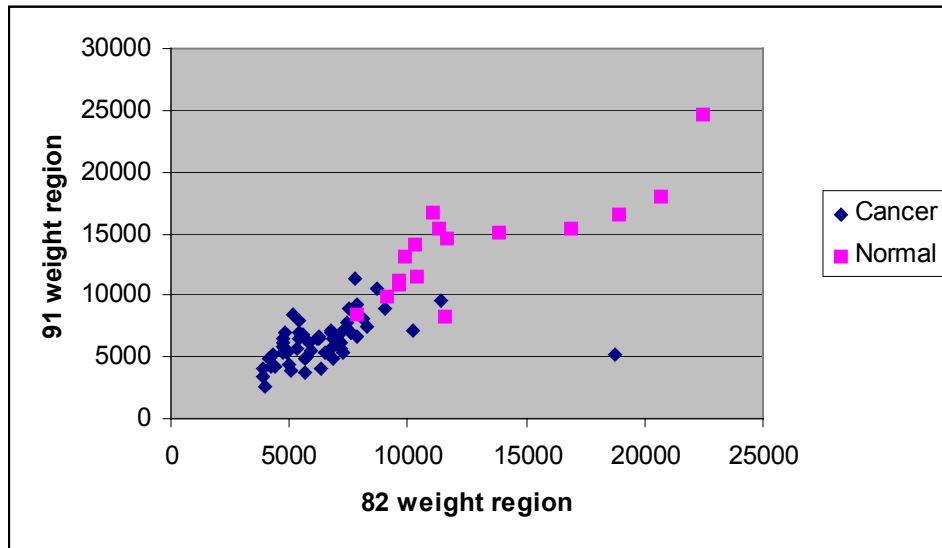


Fig. 2. 91-weight region intensity versus 82-weight region intensity of AFLP data of cancer and normal samples.

Result

Fig. 1 shows the profiles of the cancer and normal AFLP sample data. The y-axis represents the Electrophoresis peak intensity while the x-axis represents the molecular weight of the amplified fragments. The major deviations in AFLP intensities between the cancer data and the normal data are in the molecular weight regions of 82 and 91 as seen in Fig. 1. There is a significant reduction in the electrophoresis peak intensity in the two molecular weight regions in the cancer data. Fig. 2 shows the graph generated by plotting the 82-weight region intensity against the 91-weight region intensity. From the graph, the cancer and normal data set could be separated into two distinct groups. Based on this observation, a simple linear separator should be able to separate the 2 sets of data in the original input space without any mapping to the feature space. The observation was confirmed in the results obtained using a linear kernel as shown in Table 2. Of the 58 cancer samples and 16 normal samples, applying SVM with a linear kernel yields a sensitivity of 0.98 and specificity of 0.75, with 57 true positives, 12 true negatives, 4 false positives and 1 false negative.

Table 2. Calculation of Sensitivity and Specificity for linear kernel

		Disease	
		+	-
Test Results	+	57 (TP)	4 (FP)
	-	1 (FN)	12 (TN)

TP = number of true positive

FP = number of false positive

FN = number of false negative

TN = number of true negative

Sensitivity = $TP / (TP + FN)$
= 0.98

Specificity = $TN / (TN + FP)$
= 0.75

Conclusion

In conclusion, a novel approach of gastric cancer diagnosis using SVM clustering of data generated by AFLP screening method is presented. The result in this paper shows that SVM is capable of clustering the AFLP data successfully for gastric cancer detection. As AFLP is a relatively low cost procedure, requiring minimum prior sequence knowledge and biological material, there is a strong potential for SVM clustering of AFLP screening data to be used as a diagnostic tool for gastric cancer.

References

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X. et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature (Lond.)* 403: 503 -- 511.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendeix, M., M. Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature (Lond.)* 406: 536 -- 540.
- Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D, 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Nat Acad Sci USA* 97: 262-267.
- Yu-Dong Cai , Xiao-Jun Liu, Xue-biao Xu and Guo-Ping Zhou , 2001, Support Vector Machines for predicting protein structural class *BMC Bioinformatics* 2:3.
- Crookes PF, 2002. Gastric cancer. *Clinical Obstetrics & Gynecology.* 45(3):892-903, 2002 Sep.
- Ferlay, J., Bray, F., Pisani, P., and Parkin, D.M. 2001. *GLOBOCAN 2000: Cancer incidence, mortality and prevalence worldwide.* IARC Press, Lyon.

Fukao, A., Tsubono, Y., Tsuji, I., Hisamichi, S., Sugahara, N., Takano, A. 1995. The evaluation of screening for gastric cancer in Miyagi Prefecture, Japan: a population-based case-control study. *Int. J. Cancer* 60(1): 45 -- 48.

Terrence S. Furey, Nigel Duffy, Nello Cristianini, David Bednarski, Michel Schummer, David Haussler, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906-914.

Joachims, T., 1999a. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (Eds.), *Advances in Kernel Methods-Support Vector Learning*. MIT Press, Cambridge, MA.

Joachims, T., 1999b. *Proceedings of the International Conference on Machine Learning*.

Jonas, D., Meyer, H.W., Matthes, P., Hartung, D., Jahn, B., Daschner, F. D., and Jansen, B. 2000. Comparative Evaluation of Three Different Genotyping Methods for Investigation of Nosocomial Outbreaks of Legionnaires' Disease in Hospitals. *Journal of Clinical Microbiology* 38(6): 2284 - 2291.

Pisani, P., Oliver, W.E., Parkin, D.M., Alvarez, N., Vivas, J. 1994. Case-control study of gastric cancer screening in Venezuela. *Br. J. Cancer* 69(6): 1102 -- 1105.

Prochazka, M., Walder, K., and Xia, J. 2001. AFLP fingerprinting of the human genome. *Hum. Genet.* 108: 59 -65.

Sonnenburg, S., Rätsch, G., Jagota, A., Müller, K.-R., 2002. New Methods for Splice Site Recognition, *Proceedings of the International Conference on Artificial Neural Networks*.

Sujun Hua Zhirong Sun, 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721-728.

Vapnik, V., 1995. *The Nature of Statistical Learning Theory*. Springer, Berlin.

Vapnik, V., 1998. *Statistical Learning Theory*. Wiley-Interscience, New York.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407 - 4414.

Williams, K.J., Smyl, L., Lichon, A., Wong, K.Y., and Wallwork, H. 2001. Development and use of an assay based on the polymerase chain reaction that differentiates the pathogens causing spot form and net form of net blotch of barley. *Australasian Plant Pathology* 30: 37-44.

Wu CW et al, 2002. Clinical implications of chromosomal abnormalities in gastric adenocarcinomas. *Genes, Chromosomes & Cancer*. 35(3):219-31, 2002 Nov.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T., and Muller, K.-R., 2000. Engineering support vector machine kernels that recognize translation initiation sites, *BioInformatics*, 16(9):799-807.