Support Vector Machine-based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes

Li-Yeh Chuang, Kuo-Chuan Wu, Hsueh-Wei Chang, and Cheng-Hong Yang, *Member, IAENG*

***Abstract*—Oral cancer is the sixth most common cancer and a major health problem in the world. We aimed at DNA repair genes such as X-ray repair cross-complementing group (XRCC)1, 2, 3, and 4. Single nucleotide polymorphisms (SNPs) dataset with 238 samples of oral cancer and control patients were chosen for disease prediction. All prediction experiments were conducted using the support vector machine. The result showed the performances of the holdout cross validation is superior to 10-fold cross validation, and the best classification accuracy is 64.2%. Although only four SNPs were used in this analysis, our proposed methodology is still high-throughput for genome-wide SNPs. Once more SNPs were introduced to oral cancer prediction, the prediction rate will be further improved.**

***Index Terms*—oral cancer, X-ray repair cross-complementing group, single nucleotide polymorphisms, support vector machine**

I. Introduction

ORAL cancer (OC) is the sixth most common cancer and a major health problem in the world. It has been identified as a huge threat to public health because of its high morbidity and mortality. Recent researches have shown that rate of OC have been steadily diminishing among males, but have grown sharply among females. However, it is one of the fastest increasing malignancies in Taiwan. OC's occurrence is associated with exposure to smoking and alcohol consumption. Whatever, the majority of cases occur is mainly associated with betel quid chewing in Asia. In addition to genetic differences, the other risk factors include age, human papilloma virus infection, and race etc. Therefore, it is urgent for local researchers to understand the causes behind the trend in Taiwan [1-5]. With the completion of the Human Genome Project (HGP), new opportunities and challenges had been presented for uncovering the genetic basis of complex diseases via genome-wide association studies. The gold of the Human Genome Project [6] provide a tool to help scientists understanding human genetic map and to decipher the genetic code. One of the major goals of the post-genome era is to understand the role of genetics in human health and disease. After the completion of the human genome project, increasing attention has focused on the identification of human genomic variations, especially single nucleotide polymorphisms (SNPs) [7-9]. DNA damage is the most important factor for carcinogenesis because of the insults of environmental carcinogens. Repair of DNA damage can protect cells against carcinogenesis, and the polymorphisms of the DNA repair gene have been implicated as susceptibility factors in cancer development [10]. Over 130 genes coding for proteins of the various DNA repair pathways have been identified [11] and in excess of 400 SNPs characterized within the 80 genes had been screened for variation [11]. SNPs are known to be the most common variant in the human genome and play an important role for drug development, cancer and genetic disease research. SNPs are defined as single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals, these occur at appreciable frequent has an abundance of 1% or greater in the human population. With the genome-wide SNP discovery, many genome-wide association studies are likely to identify multiple genetic variants that are associated with complicated diseases [12, 13].

Machine learning tasks are applied many wide bioinformatics research. In classification or regression, which is to predict the outcome associated with a whole samples. The purpose of classification is to build an efficient model for predicting the class membership of data. Many classification approaches have been proposed, such as: Nearest-Neighbor (NN), Naïve Bayes (NB), Random forest (RF), Support vector machine (SVM) and so forth. This model should produce a correct label on the training data and predict the label of any unknown data accurately. These methods had been applied in many fields, such as: decisions involving judgment, screening images, load forecasting, marketing and sales and diagnosis [14].

In this study, we have focused our aim at X-ray repair cross-complementing group (XRCC) as it is not likely that every possible candidate genes can be investigated in a single study. We collected 238 samples of OC that applies machine learning for disease prediction using SNP data. All experiments were conducted using the Weka [14] machine learning software package with its standard settings.

L.Y. Chuang is with the Department of Chemical Engineering, I-Shou University , 84001 , Kaohsiung, Taiwan (E-mail: chuang@isu.edu.tw).

K.C. Wu is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung, Taiwan (E-mail: kuo.chuan.wu@gmail.com).

H.W. Chang is with the Department of Biomedical Science and Environmental Biology, Cancer Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, 80708, Taiwan (changhw@kmu.edu.tw).

C.H. Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung Taiwan (phone: 886-7-3814526#5639; E-mail: chyang@cc.kuas.edu.tw). He is also with the Network Systems Department, Toko University, 61363, Chiayi, Taiwan. (E-mail: chyang@cc.kuas.edu.tw).

II. Materials and methods

A. Subjects

This dataset was collected from our previous result [15] shown as TABLE I. We divided this dataset into oral cancer and control groups without considering the personal information. The type of SNP genotype is symbol, we convert to numerical as {-1, 0, 1}. For example, the genotype of SNP no.1 is CC = -1, CT = 0 and TT = 1.

TABLE I. The SNPs genotype information

SNP no.	*Gene (SNP)*	*Genotype*		
		1	*2*	*3*
1	XRCC1 (rs1799782)	CC	CT	TT
2	XRCC2 (rs2040639)	AA	AG	GG
3	XRCC3 (rs861539)	CC	CT	TT
4	XRCC4 (rs2075685)	TT	TG	GG

B. Support vector machine

SVM is the one of state-of-the-art supervised learning approach which is the prediction problem, whose goal is to build a model from a set of positively and negatively labeled training vectors that can classify unlabelled test samples. SVM establishes a maximum margin that can find the best hyperplane to separate the two categories in Euclidean space [14, 16, 17]. SVM was trained to distinguish between case and control in SNP samples. In this study, the Weka [14] was used to perform the SVM work, using the sequential minimal optimization (SMO) algorithm and radial basis function (RBF) kernel. The RBF kernel function is defined as:

$$K(X_i - X_j) = \exp(-\gamma \| X_i - X_j \|^2) \qquad (1)$$

where X_i and X_j are two feature vectors, and γ is the training parameter . The parameters are used as default in Weka. We considered OC cases as positive samples and controls as negative samples, and used SNP genotypes as categorical features. We adopted SVM to discriminate OC cases against controls in this research.

C. Cross-Validation Test

In data mining like classification problem, a typical task is to construct a model from available data such a model may be a classifier. We can't make sure that a model can predict future unseen data well, so the model needs to demonstrate the prediction capability. In statistics, a cross validation is an approach to estimate the generalization performance of prediction. Two or more learning algorithms would be compared through cross validation that can be used in a statistical hypothesis test to know that one approach is superior to another. There are three common cross validation methods including holdout cross validation (holdout CV) and *m*-fold cross validation (*m*-fold CV) shown as fallow [14, 18]:

Holdout cross validation

A simplest kind of cross validation method is called holdout cross validation that is to separate the available data into two non-overlapped sets (i.e. training set and testing set). It is common to split 2/3 of the data as the training set and the remaining 1/3 as the test set. The model maybe a classifier fits a function using the training set. And then the testing set used to predict the output for the data using the model [18].

m-fold cross validation

An improved cross validation approach from holdout validation method is called *m*-fold cross validation. In *m*-fold cross validation, the available data are separated into *m* non-overlapped and equally sized set. A variant of separated sets are randomly divide the data into training and testing sets *m* different times. The holdout method is repeated *m* times. One of the *m* subsets is used as the testing sets and the remaining *m*-1 subsets as the training sets. Then the average accuracy across all *m* trials is calculated [18].

D. Accuracy estimation

The common estimation is used in medical diagnosis including: Positive hit rate (i.e. Sensitivity, SN), Negative hit rate (i.e. Specificity, SP) and Accuracy (ACC) rate. If the case with the "positive" class (with disease) correctly classified as positive called True Positive (*TP*), however the case with the "positive" class classified as negative called False Negative (*FN*). Conversely, the case with the "negative" class (without disease) correctly classified as negative called True Negative (*TF*), while the case with the "negative" class classified as positive called False Positive (*FP*). SN is the proportion of cases with positive class that are classified as positive. On the other hand, the SP is the proportion of cases with negative class that are classified as negative. The sensitivity and specificity are computed as formula (2) and (3). The ACC rate is calculated as formula (4).

$$SN = \frac{TP}{TP + FN} \qquad (2)$$

$$SP = \frac{TN}{TN + FP} \qquad (3)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \qquad (4)$$

III. Results and discussion

A. Experimental results

Association between individual polymorphisms for DNA repair genes and oral cancer

TABLE II shows the estimated effect (odds ratio and 95% CI) of individual SNPs of XRCC1, XRCC2, XRCC3, and XRCC4 genes on the occurrence of oral cancer. Except for XRCC3 rs861539-CT, no specific SNP was significantly associated with the risk of occurrence of oral cancer. All these genotype frequencies were in agreement with the Hardy-Weinberg equilibrium.

Classifier estimation

This study performed SVM training and test analysis to observe the combination of SNPs. The prediction outcomes were estimated by the SVM classifier that distinguished the case and control SNP genotype data. The holdout cross-validation and 10-fold cross-validation were used to assess the performance of the SVM models and reported in TABLE III. We compare classifier to each SNP combination that is evaluated using cross validation. The result show, the performances of 10-fold cross validation are the same in each combination of SNPs, and the best ACC is 57.14%. The holdout cross validation is superior to 10-fold cross validation, and the best ACC is 64.2%. The prediction accuracy is underperformed using SVM. Under observation, we found the distribution of case/control of SNP data are closely, so it is difficult to classify clearly.

B. *Discussion*

Typically, the complex diseases are caused by joint factors of multiple genetic variations instead of a single genetic variation [17]. In this study, our rationale for exploring the gene–gene interactions is justified because interactions of multiple genes are widely hypothesized to influence risk for OC. The development of new analytical methods makes it feasible to systematically explore genome-wide interactions. We introduced this idea to examine the important role of combinational SNPs for four DNA repair genes XRCCs 1-4 in oral cancer. Other SNPs in different DNA repair genes that may be involved in the association of oral cancer were not included completely in this study. Our main focus was to understand the contribution to oral cancer risk of functionally relevant joint effect for combinational SNPs within and between different cancer pathways like XRCCs 1-4. There has been increasing evidence regarding the combined effect of commonly occurring SNPs on cancer risk, supported by polygenic models in cancers of the lung etc [15].

Recently, there are many computational methods have been proposed to analysis SNP data using Multifactor Dimensional Reduction (MDR), or machine learning algorithms. The gene-gene interactions of SNPs are very important in determining individual susceptibility to complex diseases [16]. Therefore, this study introduces a machine learning approach for SNP data of OC. In classification problems, overfitting appears when computationally intensive search algorithms are used. Estimates may be overfitted and yield biased predictions under these circumstances [19]. If the training data lies too closely together, the classifier predictions are of poor quality. This occurs when there is insufficient data to train the classifier and the data does not fully cover the concept being learned. This problem is common in many real world samples where the available data may be rather noisy [20]. In order to avoid overfitting, some additional techniques have been discussed, such as cross-validation, regularization, and early termination or resampling [21, 22]. We try whole possible of SNP combination for SVM that used holdout cross validation and 10-fold cross validation. In the holdout cross validation, the best accuracy of combination SNP all include XRCC3 (see TABLE III). On the other hand, because of case and control data distributions are closely and 10-fold cross validation approach increase the training set that lead to data more closely. It makes the performance is not significant in each combination SNP. The effects of SNP-SNP interaction are recognized for the biological issues previously, such as oral cancer [15], osteoporosis [23] and type 2 diabetes mellitus [16] etc. We may need more exact features (i.e. SNPs), more reliable samples and more powerful computational approach for precise disease prediction.

IV. Conclusion

The DNA repair pathways investigated in this study have been reported in oral cancer development. However, their genetic interactions, detected through variant alleles (SNPs), have not been described previously. The gold of this study to use SNP data for experiments that is to demonstrate our proposed approach can obtained advantage ability of prediction and SNP selection. Experimental results show that SVM obtained 64.2% classification accuracy. The novelty of our study is the demonstration of significant joint effect between SNPs that did not have an individual effect on oral cancer risk. Therefore, this approach for joint effect of combinational SNPs has the potential to assist in the identification of complex biological relationships among cancer processes during the development of oral cancer. In the future work, to use machine learning methods to predict other disease or to acquire more SNP of OC search for significant or helpful information.

References

[1] A. Bunnell, N. Pettit, N. Reddout, K. Sharma, S. O'Malley, M. Chino, and K. Kingsley, "Analysis of primary risk factors for oral cancer from select US states with increasing rates," Tobacco Induced Diseases, vol. 8, p. 5, 2010.

[2] S. Han, Y. Chen, X. Ge, M. Zhang, J. Wang, Q. Zhao, J. He, and Z. Wang, "Epidemiology and cost analysis for patients with oral cancer in a university hospital in China," BMC Public Health, vol. 10, p. 196, 2010.

[3] R. Mishra, "Glycogen synthase kinase 3 beta: can it be a target for oral cancer," Molecular Cancer, vol. 9, p. 144, 2010.

[4] M. Rahman, N. Ingole, D. Roblyer, V. Stepanek, R. Richards-Kortum, A. Gillenwater, S. Shastri, and P. Chaturvedi, "Evaluation of a low-cost, portable imaging system for early detection of oral cancer," Head & Neck Oncology, vol. 2, p. 10, 2010.

[5] C.-C. Su, Y.-Y. Lin, T.-K. Chang, C.-T. Chiang, J.-A. Chung, Y.-Y. Hsu, and I.-B. Lian, "Incidence of oral cancer in relation to nickel and arsenic concentrations in farm soils of patients' residential areas in Taiwan," BMC Public Health, vol. 10, p. 67, 2010.

[6] J. Watson, "The human genome project: past, present, and future," Science, vol. 248, pp. 44-49, April 6, 1990 1990.

[7] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, t. m. o. t. DOE, and NIH planning groups, "New Goals for the U.S. Human Genome Project: 1998-2003," Science, vol. 282, pp. 682-689, October 23, 1998 1998.

[8] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," BMC bioinformatics, vol. 10, p. S65, 2009.

[9] A. V. Kulkarni, N. S. Williams, Y. Lian, J. D. Wren, D. Mittelman, A. Pertsemlidis, and H. R. Garner, "ARROGANT: an application to manipulate large gene collections," Bioinformatics, vol. 18, pp. 1410-1417, November 1, 2002 2002.

[10] E. L. Goode, C. M. Ulrich, and J. D. Potter, "Polymorphisms in DNA Repair Genes and Associations with Cancer Risk," Cancer Epidemiology Biomarkers & Prevention, vol. 11, pp. 1513-1530, December 1, 2002 2002.

[11] H. W. Mohrenweiser, D. M. Wilson, and I. M. Jones, "Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes," Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, vol. 526, pp. 93-125, 2003.

[12] S. Buch, C. Schafmayer, H. Völzke, C. Becker, A. Franke, H. v. Eller-Eberstein, C. Kluck, I. Bässmann, M. Brosch, F. Lammert, J. F.

Miquel, F. Nervi, M. Wittig, D. Rosskopf, B. Timm, C. Höll, M. Seeger, A. ElSharawy, T. Lu, J. Egberts, F. Fändrich, U. R. Fölsch, M. Krawczak, S. Schreiber, P. Nürnberg, J. Tepel, and J. Hampe, "A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease," Nature genetics, vol. 39, pp. 995-999, 2007.

[13] B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Tenesa, S. M. Farrington, J. Prendergast, S. Olschwang, T. Chiang, E. Crowdy, V. Ferretti, P. Laflamme, S. Sundararajan, S. Roumy, J.-F. Olivier, F. Robidoux, R. Sladek, A. Montpetit, P. Campbell, S. Bezieau, A. M. O'Shea, G. Zogopoulos, M. Cotterchio, P. Newcomb, J. McLaughlin, B. Younghusband, R. Green, J. Green, M. E. M. Porteous, H. Campbell, H. Blanche, M. Sahbatou, E. Tubacher, C. Bonaiti-Pellié, B. Buecher, E. Riboli, S. Kury, S. J. Chanock, J. Potter, G. Thomas, S. Gallinger, T. J. Hudson, and M. G. Dunlop, "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24," Nature genetics, vol. 39, pp. 989-994, 2007.

[14] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2 ed. San Francisco: Morgan Kaufmann, 2005.

[15] C.-Y. Yen, S.-Y. Liu, C.-H. Chen, H.-F. Tseng, L.-Y. Chuang, C.-H. Yang, Y.-C. Lin, C.-H. Wen, W.-F. Chiang, C.-H. Ho, H.-C. Chen, S.-T. Wang, C.-W. Lin, and H.-W. Chang, "Combinational polymorphisms of four DNA repair genes XRCC1, XRCC2, XRCC3, and XRCC4 and their association with oral cancer in Taiwan," Journal of Oral Pathology & Medicine, vol. 37, pp. 271-277, 2008.

[16] H.-J. Ban, J. Y. Heo, K.-S. Oh, and K.-J. Park, "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine," BMC Genetics, vol. 11, p. 26, 2010.

[17] R. V. Spriggs, Y. Murakami, H. Nakamura, and S. Jones, "Protein function annotation from sequence: prediction of residues interacting with RNA," Bioinformatics, vol. 25, pp. 1492-1497, June 15, 2009 2009.

[18] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, pp. 273-324, 1997.

[19] J. Reunanen, I. Guyon, and A. Elisseeff, "Overfitting in Making Comparisons Between Variable Selection Methods " Journal of Machine Learning Research, vol. 3, 2003.

[20] J. Loughrey and P. Cunningham, "Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets," in Research and Development in Intelligent Systems XXI, ed, 2005, pp. 33-43.

[21] C. Schaffer, "Overfitting avoidance as bias," Machine learning, vol. 10, pp. 153-178, 1993.

[22] D. H. Wolpert, "On overfitting avoidance as bias," Technical Report SFI-TR-92-03-5001, Santa Fe Institute1993.

[23] H.-W. Chang, L.-Y. Chuang, C.-H. Ho, P.-L. Chang, and C.-H. Yang, "Odds Ratio-Based Genetic Algorithms for Generating SNP Barcodes of Genotypes to Predict Disease Susceptibility," OMICS: A Journal of Integrative Biology, vol. 12, pp. 71-81, 2008.

TABLE II. Estimated effect (odds ratio and 95% CI) of individual SNP of XRCC1, XRCC2, XRCC3, and XRCC4 genes on the occurrence of oral cancer

Gene	*SNP genotype*	*Number of control / number of cases*	*Crude odds ratio*	*95% CI*	*p-value*
XRCC1 (rs1799782)	1:CC	23/17	1.00		
	2:CT	86/83	1.31	0.65-2.62	0.45
	3:TT	19/10	0.71	0.26-1.92	0.50
XRCC2 (rs2040639)	1:AA	20/18	1.00		
	2:AG	75/48	0.71	0.34-1.48	0.36
	3:GG	33/44	1.48	0.68-3.23	0.32
XRCC3 (rs861539)	1:CC	122/96	1.00		
	2:CT	6/14	2.97	1.10-8.00	0.03
	3:TT	0/0			
XRCC4 (rs2075685)	1:TT	4/4	1.00		
	2:TG	47/39	0.83	0.19-3.54	1.00
	3:GG	77/67	0.87	0.21-3.61	1.00

XRCC, X-ray repair cross-complementing group; SNP, single nucleotide polymorphism; CI, confidence interval.

TABLE III. Performance measures for the holdout cross validation and 10-fold cross validation of the SVM to predict each combination of SNPs data of oral cancer.

Holdout cross validation			**10-fold cross validation**			
SN	**SP**	**ACC**	**SN**	**SP**	**ACC**	**SNP selected**
0.114286	0.73913	0.469136	0	1	0.537815	XRCC 1
0	1	0.567901	0	1	0.537815	XRCC 2
0.228571	**0.956522**	**0.64198**	0.127273	0.953125	0.571429	XRCC 3
0	1	0.567901	0	1	0.537815	XRCC 4
0.114286	0.73913	0.469136	0	1	0.537815	XRCC 1,2
0.285714	0.717391	0.530864	0.127273	0.953125	0.571429	XRCC 1,3
0.114286	0.76087	0.481481	0	1	0.537815	XRCC 1,4
0.228571	**0.956522**	**0.64198**	0.127273	0.953125	0.571429	XRCC 2,3
0	1	0.567901	0	1	0.537815	XRCC 2,4
0.228571	**0.956522**	**0.64198**	0.127273	0.953125	0.571429	XRCC 3,4
0.285714	0.717391	0.530864	0.127273	0.953125	0.571429	XRCC 1,2,3
0.114286	0.76087	0.481481	0	1	0.537815	XRCC 1,2,4
0.285714	0.717391	0.530864	0.127273	0.953125	0.571429	XRCC 1,3,4
0.228571	**0.956522**	**0.64198**	0.127273	0.953125	0.571429	XRCC 2,3,4
0.257143	0.782609	0.555556	0.127273	0.953125	0.571429	XRCC 1,2,3,4