Identifying FCN2 as a Potential Biomarker for Hepatocellular Carcinoma Using a Novel Supervised Classification Method

Nuttachat Wisittipanit*, Ekachai Chukeatirote

School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand

**Corresponding author: nuttachat.wis@mfu.ac.th*

ABSTRACT

Hepatocellular carcinoma (HCC) is the most prevalent form of liver cancer of which most patients show no symptoms of the condition until the ailment is beyond the curable stage. Computational analysis using data mining methods were employed to analyze gene expression profiles on microarray chips from 511 samples (268 HCC tissues and 243 adjacent non-HCC tissues) to identify highly differentially expressed genes. Of 23,000 genes available in the platform, 500 genes (~2%) were preliminary identified by Relief algorithm based on the high ranking scores of their expression levels between HCC tissues and normal tissues. Then, the intensive search for differentially expressed genes were performed by machine learning methods; namely Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Naïve Bayes Classification (NBC) algorithms, by attributing pool sweep process. Results showed that several highly differentially expressed genes were detected having relatively high classification accuracy with Ficolin-2 acting as a common detected gene.

Keywords: hepatocellular carcinoma; gene expression; machine learning; attribute selection, genetic marker

INTRODUCTION

Liver is the biggest and one of the most essential organs in our body. This necessary organ collects nutrients (mainly fat and protein), filters toxin out of circular blood flow, synthesizes albumin which regulates osmotic pressure in cells and tissues, and transfer important hormones to other organs (Maton, 1993). Thus, if the liver could not function properly due to conditions such as liver failure, cirrhosis or cancer, it could have terrible effects on the body and would likely lead to death (Baffy *et al*., 2011). In the case of liver cancer, the majority of patients often have no awareness of their conditions since no symptoms are apparent at the beginning stage. Therefore, when the cancer-related symptoms do emerge, discerning the patients, the cancer is often in the late stage and already deteriorates their health conditions severely such that their chance to survive is close to being non-existent (Zakim and Boyer, 2003). Liver cancer is found in population around the world especially in Asia and Africa (Jemal *et al*., 2011). There are four types of liver cancer: (i) angiosarcomas (ii) cholangiocarcinoma (iii) hepatoblastoma and (iv) hepatocellular carcinoma (HCC). Among all types of liver cancer, HCC is the most prevalent form.

Currently, the diagnostic of HCC is mostly done by measuring alpha-fetoprotein level in the blood and sometimes together with ultrasound or CT scan method (to visualize the malignant tumor) (Anand *et al*., 2008). However, problems with measuring the level of alpha-fetoprotein as an HCC diagnostic is that its sensitivity and specificity are relatively low (Lamb *et al*., 2011). Moreover, the ultrasound or CT scan are quite expensive such that patients frequently refuse (Barghini *et al*., 2013).

There are copious studies aiming to scrutinize the causes and mechanism of HCC in order to find a new HCC biomarker with higher sensitivity and specificity than the alpha-fetoprotein measurement (Patil *et al*., 2013). The widespread method at present is the microarray technology – a practice that gauges gene expression levels in human cells (Maass *et al*., 2010). Most often, the technique is employed to find differentially expressed genes between normal and malignant cells and then some statistical methods are used to identify those genes. Such genes might have the potential to reveal the cause or mechanism of HCC. A study by Wang and his team (Wang *et al*., 2009) revealed that *PEG10* gene, responsible for apoptosis inhibition of HCC cells via collaboration with *SIAH1* gene (apoptosis mediator), was detected to be highly expressed in HCC tissues. However, this study only acquired a small number of HCC and adjacent non-HCC tissues and might not be enough for accurate prediction of differentially expressed genes; additionally, *FLJ10743, CRP and EGR1* were examples of down-regulated genes in HCC tissues detected in the study. Moreover, significant analysis and gene set enrichment

Received 28 October 2017; Accepted 22 July 2019

analysis were used to determine differentially expressed genes in HCC tissues (Skawran *et al*., 2008) that found *SCAMP3, IQGAP3, PYGO2, GPATC4* and *ASH1L* genes to be highly expressed. Machine learning algorithms were employed in some studies to discover knowledge or patterns in gene expression profiles from the microarray technology. For examples, differences in gene expressions between primary metastasis-free HCC and primary HCC with metastasis were determined using Compound Co-variate Predictor (CCP) where a univariately significant group of genes were obtained and the leading gene, osteopontin, was shown to be greatly expressed in metastatic HCC (Ye *et al*., 2003); the CCP classifier had a performance of ~93% in accuracy classifying between two different types of HCC tumour. Research by Jia and colleague (Jia *et al*., 2007) used Nearest Shrunken Centroid (NSC) algorithm to classify independent HCC and non-tumour samples on microarray data with only five selected genes (*GPC3*, *PEG10*, *MDK*, *SERPINI1* and *QP-C*); the performance of NSC algorithm was ~85.5% in accuracy. Moreover, Support Vector Machine (SVM) was utilized as a classification method between healthy control and HCC samples on microarray data with a few groups of genes clustered by network topology (Shen and Liu, 2017); the SVM classifier's performance was ~0.88 AUC (Area Under the Curve), averaged among four significant clusters of genes.

The purpose of this research was to analyze gene expression data gathered from a microarray experiment performed on human liver tissues (HCC and non-HCC) by using data mining and machine learning methods that intended to select differentially expressed genes between HCC tissues and adjacent non-HCC tissues. The processes involved feature selection method and feature sweeping procedure done by three machine learning algorithms: NBC, KNN and SVM. The results indicated that the gene expression profiles from tissues in HCC and adjacent non-HCC can be well classified by those algorithms with relatively high accuracy and *FCN2* was a common gene found by all three machine learning algorithms.

MATERIALS AND METHODS

The data used in this study were gene expression levels determined by microarray. The gene expression data were obtained from 268 Asian patients comprising 511 samples and 37,582 genes, 268 samples from HCC tissues (from 268 patients) and 243 from adjacent non-HCC tissues (from 243 patients), were investigated. There are a total of 243 matched HCC tumor and adjacent non-HCC samples with 25 HCC tissues having no non-HCC pair samples available. Those data sets were available at Gene Expression Omnibus database with identification number GSE25097 (Tung *et al*., 2011).

Data normalization

For each gene, the intensity readings from the microarray data were tweaked to have values ranging from -1 to 1. To achieve this normalization, the lowest intensity value of each attribute was set to -1 and the maximum intensity value was set to 1. Then, the other intensity values were scaled within that range. Data normalization avoids the condition that attributes in higher value ranges could dominate those in lower value ranges and also prevent arithmetical complexities faced during the computation.

Initial attribute selection: Relief algorithm

The attribute selection was conducted to search for significant genes or attributes that are more likely to distinguish the HCC tissue samples from the adjacent-non HCC tissue samples in order to vastly improve the classifier performances. The Relief Algorithm (Kira and Rendell, 1992) was used to select genes based on the notion that the intensity values of distinctive class samples should be dissimilar and intensity values of the same class samples should be similar. Genes more likely to suit that notion are ranked higher and more plausible to be selected; those genes have larger difference in expression values between HCC and adjacent-non HCC tissues than lower ranked genes.

Supervised classification methods

Given a training dataset containing a class of positive instances S^+ and a class of negative instances S^-. A classification algorithm learns the pattern in the training dataset associated with either positive or negative classes (training dataset) and then form a discriminant function used to predict which class an unseen instance should belong to. Normally, a dataset is divided into 2 datasets: a training dataset and a test dataset in order to train a classifier, choose appropriate classifier's parameters and test accuracy of the classifier. The dataset division were done multiple times to obtain optimal parameters. Three classification algorithms were used in this study: (i) Naïve Bayes (NB), (ii) K-Nearest Neighbour (KNN) and (iii) Support Vector Machine (SVM). NB classification algorithm (Domingos and Pazzani, 1997) is a probability-based classification technique which uses Bayes' theorem with solid independence hypothesis. The conditional

probabilities of dependent variables from the training dataset are approximated. The probabilities are then used for classifying new data instances. The algorithm is suitable for dataset with many discrete attributes; however, it is quite time-consuming to classify datasets with several continuous attributes. KNN (Skawran *et al*., 2008) is a classification algorithm which classifies instances to classes based on the nearest instances in the attribute space using a distance function. KNN algorithm has one essential parameter, k, i.e. when given an instance S_i of an unknown class, the KNN algorithm selects k nearest instances based on the training dataset. The class of S_i is chosen using the majority vote of k nearest neighbors employing a distance function (Maass, 2010). In other words, S_i is assigned to a class which is the majority of *k* instances. In this study, the Euclidian distance was selected as the distance function that evaluates an unknown instance with the training instances. The classification accuracy of KNN algorithm is mainly controlled by the k parameter and the choice of the distance function.

Lastly, the SVM algorithm defines a classifying function $f(X)$ as follows,

$$f(X) = \sum_{X_i \in S^+} \lambda_i^+ K(X, X_i) - \sum_{X_i \in S^-} \lambda_i^- K(X, X_i), \quad (1)$$

where λ_i^+ and λ_i^- are non-negative parameters calculated during a classifier training in which a quadratic objective function is maximized. $K(.,.)$ is the kernel function computed during the training as well. A penalty factor, C, is presented during the training step and acts as the ratio of the classification errors. Given Equation 1, a new instance S^i is assigned to a positive or negative class depending on whether the value of $f(X)$ is positive or negative. Additionally, the positive or negative values of $f(X)$ can signify the classification strength of the instances (Kilpatrick *et al*., 2003; Siegel *et al*., 2014,). The radial basis function (RBF) was utilized in this study which was defined by

$$K(X, Y) = \exp\left(-\gamma \left|\left|X - Y\right|\right|^2\right), \gamma > 0\,. \quad (2)$$

For the SVM algorithm used in this study, there are 2 parameters; C *and* γ to be adjusted to find the optimal parameters (Boser *et al*., 1992).

Computational Methodology

The computational pipeline for the HCC gene expression data between HCC tissues and adjacent non-HCC tissues is shown in Figure 1. The pipeline consists of two major procedures; (1) Data Processing (2) Attribute Selection. The Data Processing procedure is the process where the attributes are normalized in order to reduce the computational load of the second procedure. The Attribute Selection procedure consists of two steps. The first step is the initial attribute selection process called 'Relief Algorithm' where a handful of attributes are selected according to the relevance to their own class using the Relief algorithm. The second step is the intensive attribute selection process called "Attribute Sweeping" where attributes are selected by the bottom-up screening process using classification algorithms.

To compare the performances of those distinct attribute selection techniques, classification accuracies for three cases are reported: (i) using all attributes, designated as ATS_ALL (ii) using attributes selected from Relief algorithm, designated as ATS_REF and (iii) using attributes selected from "Attribute Sweeping" process, designated as ATS_SWP.

Model Selection

As mentioned previously, the Relief algorithm was used to pick significant genes from the patient samples having the ability to likely separate the affected tissues from the non-affected. The genes were sorted from the highest to the lowest according to the scoring system by the algorithm. Then, top-ranked 500 genes out of 37,582 were chosen. And out of 500 genes, the highest ranked 50 genes are again picked to be processed further.

For the Attribute Sweeping process, a bottom-up sweeping search method was used which is essentially a one-by-one model selection technique for each classification algorithm. The process used the 50 attributes chosen by the Relief algorithm. The computational steps for the bottom-up search method can be shortly explained as follows: the initial number of attributes is one and a classifier is tested to find which attribute generates the best classification accuracy; such attribute is kept, and removed from the attribute list. In the next round, the method sets the former best attribute as the basis attribute adding one attribute from the list at a time and then tests to determine which attribute (adding to the previous base) generates the best accuracy. Such attribute is stored again in the attribute set and removed from the attribute list. The searching iteration is run until the classification accuracy stop increasing.

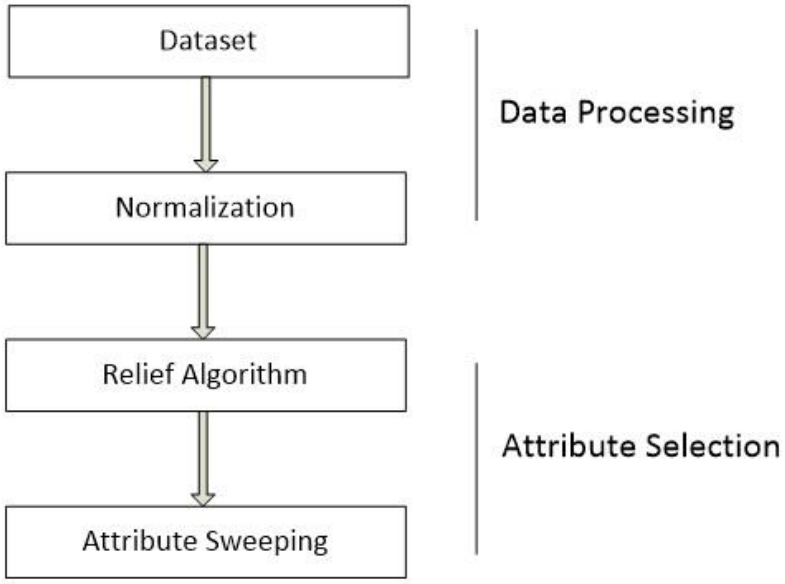

Figure 1 Computational pipeline of the novel supervised classification method with "Data Processing" and "Attribute Selection" steps.

Three classification algorithms were used in the Attribute Sweeping process; NB, SVM and KNN. For NB, there was no selection of parameter for the model selection. For KNN, there was a k parameter to be searched which fell into a range of 1 to 25. For SVM, the RBF kernel was used in which the selection of parameters C and γ was accomplished by a grid-search executing log base 2 of -5 to 15 with a step increase of 2 and log base 2 of -15 to 3 with also a step increase of 2, respectively. For each classification algorithm, five-fold cross validation was implemented, meaning that one-fifths of the samples were separated out for testing purpose and the remaining four-fifths were employed as training samples. The iteration was done five times and the classification performances were averaged. The classification performance includes classification accuracy and area under the ROC (receiver operating characteristic) curve, called "AUC". The accuracy values were calculated by dividing the number of correct predictions with the total number of predictions (one-fifths of the samples). The AUC values were calculated by integration to determine the area under the ROC curve which evaluates how efficacy the supervised-learner (classifier) is in discriminating between samples of HCC and adjacent non-HCC tissues and is also a good assessment of classification results particularly when the numbers of two classes are biased.

All the classification algorithms were from the ORANGE machine learning library found at http://www.ailab.si/orange/.

RESULTS

In this research, the gene expression data from HCC patients were analyzed by classification algorithms and attribute selection methods. The objective of this study was to investigate how efficient the computational methods were in differentiating expression profiles between HCC and non-HCC tissues and which genes were differentially expressed. Another purpose of this study was to examine which genes were shared among three classification algorithms such that those shared genes might be more significant than others. Finally, the performances of three classification algorithms used in the computational pipeline were compared.

The primary research methodology compared the performances of the three classification algorithms using three sets of genes. The first set of genes was the 500 genes of the patient samples. The second set was fixed at 50 attributes selected by Relief algorithm and the third set was chosen by the model selection method for each classification algorithm. The reason that those three sets of attributes were used for the analysis was to evaluate effectiveness of the gene selection methods in choosing a set of genes that significantly differentiate between liver tissue samples belonging to HCC and those belonging to healthy adjacent tissues by the expression level.

The hypothesis behind the computational pipeline that employed different steps of gene selection techniques was that when the number of attributes was reduced in the dataset due to subsequent gene selections, a dataset with lower number of genes would have

higher level of differences between samples of HCC and those of non-HCC tissues. The hypothesis was based on the principle that genes having more ability to differentiate HCC tissues from adjacent non-HCC tissues would more likely to be selected.

In condensed terms, there were three aspects to be investigated in this study; (i) examining whether datasets with lower number of genes would demonstrate higher differences between HCC tissues and adjacent non-HCC tissues (ii) testing which classification algorithms performed better in distinguishing the HCC tissues from healthy tissues.

Table 1 reports the results of the computational pipelines which are fundamentally the classifications between samples from HCC tissues versus those of adjacent non-HCC tissues with different set of attributes. The table demonstrates the computational pipeline results through three classification algorithms: Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), conducted through three cases of attribute selections: (i) ATS_ALL, (ii) ATS_REF and (iii) ATS_SWP. Note that the number of attributes for the case ATS_ALL is 500 and for the case ATS_REF is fixed at 50. The numbers of samples belonging to HCC and non-HCC tissues are reported as pairs of classes together with classification accuracies and AUC values.

From the results in Table 1, the classification accuracies and AUC values tended to increase when the number of genes decreased showing that datasets with lower number of genes (selected at the computational pipeline) lead to higher variances between patient samples of HCC tissues and those of adjacent non-HCC tissues. In other words, smaller sets of genes selected by Relief algorithm and model selection method could differentiate different class samples better than the full set of genes with those from model selection method appearing to distinguish the best.

For the classification performances of NB, KNN and SVM classifiers, the classification accuracy with ATS_ALL were relatively low with average accuracy values of 62.41%, 64.28% and 68.55%, respectively. In comparison, the ATS_REF for NB, KNN and SVM demonstrate the average accuracy values of 68.22%, 68.47% and 79.14%, a much improvement in accuracy across all classification algorithms. Either the ATS_ALL or ATS_REF case, SVM classifier performed the best. Lastly, the ATS_SWP for NB, KNN and SVM showed the average accuracy values of 79.86%, 78.86% and 92.74% (AUC values of 0.77, 0.74 and 0.92), respectively, which was a significant improvement from the ATS_REF case in term of accuracy measurement and AUC values. In the ATS_SWP case, the SVM classifier again performed the best. Consequently, SVM classifier appeared to perform better than NB and KNN in discriminating HCC tissue samples from adjacent non-HCC samples. Figure 2 shows classification results of three algorithms against three cases of attribute selections (ATS_ALL, ATS_REF and ATS_SWP) for accuracy (2.A) and AUC (2.B).

Table 1 Performances of three supervised classification algorithms for all three class types

Algorithm	Class Type	HCC/NHC	268/243
NBC	ATS_ALL	Accuracy (%)	62.41
		AUC	0.61
	ATS_REF	Accuracy (%)	68.22
		AUC	0.67
	ATS_SWP	Accuracy (%)	79.86
		AUC	0.77
KNN	ATS_ALL	Accuracy (%)	64.28
		AUC	0.62
	ATS_REF	Accuracy (%)	68.47
		AUC	0.67
	ATS_SWP	Accuracy (%)	78.86
		AUC	0.74
SVM	ATS_ALL	Accuracy (%)	68.55
		AUC	0.69
	ATS_REF	Accuracy (%)	79.14
		AUC	0.79
	ATS_SWP	Accuracy (%)	92.74
		AUC	0.92

HCC = HCC tissues, NHC = adjacent non-HCC tissues, AUC = Area under the curve

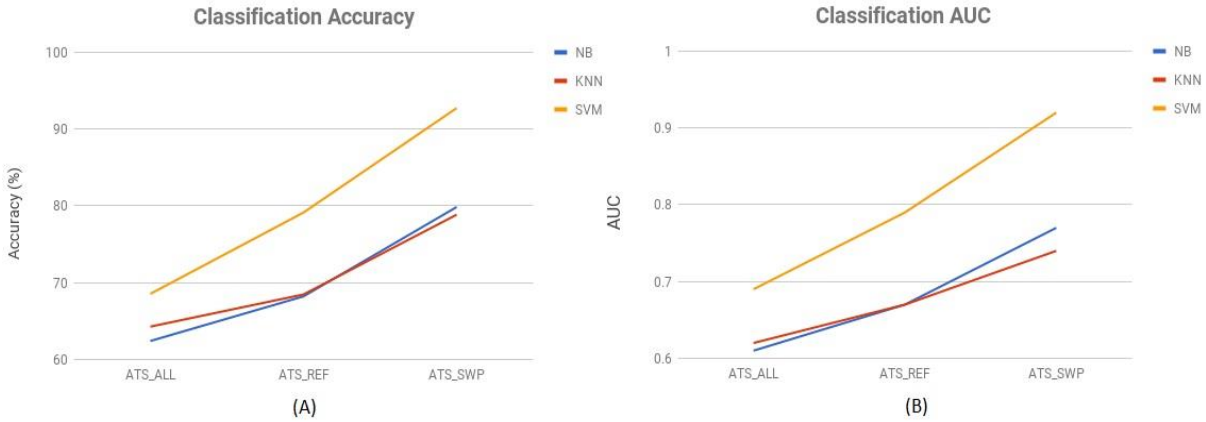

Figure 2 Classification results of NBC, KNN and SVM with attribute selection cases (ATS_ALL, ATS_REF, ATS_SWP) for accuracy (A) and AUC (B).

Table 2 shows the number of genes selected by ATS_SWP class type for NB, KNN and SVM classification algorithms which were 4, 3 and 3, respectively. Six unique genes were discovered which accounts for approximately 0.00026% of the total genes initially available at the microarray platform. Each gene in the table is represented by GenBank accession, gene name, expression side and p-value. Expression side indicates which side (HCC or adjacent non-HCC tissues) had higher level of gene expression. P-value was calculated using a standard student t-test. Two genes were common between 2 classification algorithms; *LIFR* and *MT1F* which encode leukaemia inhibitory factor receptor precursor protein and metallothionein-1F isoform 1 protein, respectively. *LIFR* product can induce the last stage differentiation of leukaemia cells and *MT1F* product has the ability to bind numerous heavy metals. There was also one common gene found among all classification algorithms which was *FCN2* gene that encodes protein Ficolin-2 functioning in immune system through the activation of lectin pathway and also helps improve phagocytosis of *Salmonella typhimurium*. Other interesting genes discovered by the algorithms include *SLCO1B3*, whose product is related in renal removal of organic anions; *CXCL14*, whose product is a ubiquitin receptor resulting in higher level of intracellular calcium ions and lower level of cellular cAMP (cyclic adenosine monophosphate) levels.

Table 2 Characteristic genes expressed in HCC vs adjacent non-HCC tissues.

Algorithm	GenBank Accession	Gene Symbol	Expression Side	P-Value
NB	NM_015837	FCN2	non-HCC	2.84E-132
	NM_005950	MT1G	non-HCC	3.17E-95
	NM_019844	SLCO1B3	non-HCC	1.71E-60
	NM_002310	LIFR	HCC	2.56E-173
KNN	NM_015837	FCN2	non-HCC	2.84E-132
	NM_005949	MT1F	non-HCC	3.34E-135
	NM_002310	LIFR	HCC	2.56E-173
SVM	NM_015837	FCN2	non-HCC	2.84E-132
	NM_005949	MT1F	non-HCC	3.34E-135
	BC003513	CXCL14	non-HCC	3.01E-136

DISCUSSION

In this study, a computational pipeline method using gene selections and classification algorithms was developed to examine the differential expressed genes of liver tissue samples between HCC and adjacent non-HCC tissues in 268 Asian patients. The gene expression level in those tissues were detected by microarray technique (Tung *et al*., 2011). Two stages of gene selections were used to select the genes that best differentiate HCC from adjacent non-HCC tissues.

The results show that HCC and adjacent non-HCC tissues were classified well across all

classification algorithms especially SVM and KNN. The increases in classification accuracies and AUC values were demonstrated when using the Relief algorithm and model selection method (in which the performances of SVM and KNN classification algorithms outperformed those of NB algorithms). While *LIFR* was commonly selected by NB and KNN algorithms and *MT1F* was commonly selected by KNN and SVM algorithms, *FCN2* was detected by all the classification algorithms used in this study; thus, this gene could play an important role in the development of HCC.

Similar studies on gene expression analysis of HCC versus non-HCC samples using machine learning methods yielded slightly less performance values i.e. accuracy and AUC when compared to those obtained by SVM in this study of ATS_SWP class type; for instances, the study that employed NSC algorithm attained about 85.5% in classification accuracy (Jia *et al*., 2007) while another study that used SVM achieved approximately 0.88 AUC (Shen and Liu, 2017) (classification accuracy was not reported in the study). However, the performance of those studies were fairly better than NBC and KNN (all class types) and SVM (all class types except ATS_SWP). Additionally, they found completely different sets of significant genes, indicating that different algorithms, when applied to similar data but under distinct circumstances and patients, can lead to totally unique results. This implies that other factors might play an essential role in HCC physiology.

The *FCN2* gene encodes ficolin-2 protein which is one of the three ficolin proteins in human and is predominantly synthesized in the liver and involved in the innate immune system through the lectin pathway as it is a pattern recognition receptor (PRR) which can identify specific molecules associated with certain pathogens (Matsushita *et al*., 2000). Thus, when the production of ficolin-2 drops, the immune performance in HCC tissues could also be reduced. Interestingly, a research in connection of hepatitis B virus and HCC to human ficolin-2 discovered that the expression levels of ficolin-2 in both serum and intrahepatic was significantly low in HCC and in cirrhosis patients compared to those in healthy control individuals (Chen *et al*., 2015). Moreover, ficolin-2 deficiency was found to be associated with haematological malignancies (Kilpatrick *et al*., 2003) – malignant tumor affecting bone marrow and lymphatic system. A study investigating a relationship between ovarian cancer and ficolin proteins (ficolin-2 and ficolin-3) found that the expression level of *FCN2* gene in patients with ovarian cancer is significantly less than that of healthy controls (Szala *et al*., 2013).

A study by Wang *et al* (Wang *et al*., 2009) that analyzed gene expression data of 22 HCC tissues and para-cancerous liver tissues showed that the *MT1G* gene was found to be differentially expressed. It is noteworthy to point out that the *MT1G* gene codes metallothionein-1G, a protein containing high level of cysteine that can bind numerous heavy metals. It was suggested that a cell lacking the expression of *MT1G* might lose the ability to export some heavy metals, such as arsenic, barium and cadmium, out of the liver (Foster *et al* 1988), which would lead to an accumulation of heavy metal in the cancer cells. The *MT1F* gene encodes metallothionein-1F that has the similar function as the protein encoded by *MT1G*. Another noteworthy gene, *SLCO1B3*, which was detected only by the Naïve Bayes algorithm, encodes a protein named Solute carrier organic anion transporter family member 1B3. This protein is a member of the solute carrier (SLC) group comprising of membrane transport proteins which are mostly situated in the cell membrane. Its duty includes facilitating the sodium-independent uptake of several compounds and, importantly, involving in the removal of organic anions and bile acids from the liver (Van De Steeg *et al*., 2012). This implies that the lower expression of *SLCO1B3* gene in HCC tissues could lead to the retention of those two substrates. Moreover, leukaemia inhibitory factor receptor alpha (LIFR) protein, which is encoded by the *LIFR* gene (the only gene that was up-regulated in the HCC tissues) commonly identified by Naïve Bayes and K-Nearest Neighbour algorithms, acts as a receptor that binds to the leukaemia inhibitory factor protein - a cytokine protein that deters differentiation of myeloid leukemic cells; subsequently, the LIFR protein helps regulates cell proliferation. The mutation in *LIFR* gene or its promoter can cause different types of tumors such as Schwartz-Jampel syndrome and salivary gland pleomorphic adenoma (Timmermann *et al*., 2002). Finally, the *CXCL14* gene, only detected by the SVM algorithm, encodes the protein C-X-C motif chemokine 14 belonging to the chemokine family which is implicated in inflammatory activities by regulating leukocyte intrusion into inflamed area. It also functions as a chemoattractant in order to direct the movement of immune cells such as T-cells and B-cells (Hromas *et al*., 1999).

With the novel supervised classification method to identify differentially expressed genes in HCC vs adjacent non-HCC tissues, numerous genes were detected as potentials for HCC biomarkers. The

gene selection pipeline proves to be effective as it helps increase the classification performance significantly. The SVM showed to be the best classifier among the three classification algorithms by providing the highest accuracy and AUC value. All of the discovered genes were expressed higher in the non-HCC samples except for the *LIFR* gene, meaning that the expressions of most detected genes were significantly lower in HCC tissues than those in adjacent non-HCC tissues. It would be interesting in further research how these genes are involved in HCC mechanism and pathway. Among those identified genes, the *FCN2* gene, was found to be substantially under express in HCC tissues and might be the most promising gene to be tested further for the HCC biomarker approval. Even though *FCN2* is regarded as the highly differential expressed gene because of its common detection among all three supervised classification method and should be investigated experimentally, other five detected genes (*MT1G, SLCO1B3, LIFR, MT1F* and *CXCL14*) might be considered as worthwhile targets for further analysis as well.

ACKNOWLEDGEMENTS

This research was supported by the Mae Fah Lung University research grant, fiscal year 2016.

REFERENCES

Anand P, Kunnumakkara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung B, Aggarwal BB (2008) Cancer is preventable disease that requires major lifestyle changes. Pharm Res 25: 2097-2116.

Baffy G, Brunt EM, Caldwell SH (2012) Hepatocellular carcinoma in non-alcoholic fatty liver disease: An emerging menace. J Hepatol 56: 1384-1391.

Barghini V, Donnini D, Uzzau A, Saordo G (2013) Hepatocellular carcinoma – future outlook. IntechOpen, London, pp 197-215.

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. Proceedings 5th Annual Conference Workshop on Computational Learning Theory, pp 144-152.

Chen T, Hu Y, Ding Q, Yu J, Wang F, Luo F, Zhang XL (2015) Serum ficolin-2 concentrations are significantly changed in patients with hepatitis B virus infection and liver diseases. Virol Sin 30: 249-260.

Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29: 103-130.

Foster R, Jahroudi N, Varshney U, Gedamu L (1988) Structure and expression of the human metallothionein-1G gene. Differential promoter activity of two linked metallothionein-1 genes in response to heavy metals. J Biol Chem 263: 1528-1535.

Gebo KA, Chander G, Jenckes MW, Ghanem KG, Herlong HF, Torbenson MS, El-Kamary SS, Bass EB (2002) Screening tests for hepatocellular carcinoma in patients with chronic hepatitis C: A systematic review. Hepatology 36: 84-92.

Hromas R, Broxmeyer HE, Kim C, Nakshatri H, Christopherson K 2nd, Azam M, Hou YH (1999) Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells. Biochem Biophys Res Commun 255(3): 703-706.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA Cancer J Clin 61: 69-90.

Jia HL, Ye QH, Qin LX, Budhu A, Forgues M, Chen Y, Liu YK, Sun HC, Wang L, Lu HZ, Shen F, Tang ZY and Wang XW (2007) Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma. Clin Cancer Res 13: 1133-1139.

Kilpatrick DC, Mclintock LA, Allan EK, Copland M, Fujita T, Jordanides NE, Koch C, Matsushita M, Shiraki H, Stewart K, *et al.* (2003) No strong relationship between mannan binding lectin or plasma ficolins and chemotherapy-related infections. Clin Exp Immunol 134: 279-284.

Kira K, Rendell LA (1992) A practical approach to feature selection. Proceedings of the 9th International Conference on Machine Learning, pp 249-256.

Lamb JR, Zhang C, Xie T, Wang K, Zhang B, Hao K, Chudin E, Fraser HB, Millstein J, Ferguson M, *et al.* (2011) Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One doi:10.1371/journal.pone.0020090.

Maass T, Sfakianakis I, Staib F, Krupp M, Galle PR, Teufel A (2010) Microarray-based gene expression analysis of hepatocellular carcinoma. Curr Genomics 11: 261-268.

Matsushita M, Endo Y, Fujita T (2000) Cutting Edge: Complement-activating complex of ficolin and mannose-binding lectin-associated serine protease. J Immunol 164: 2281-2284.

Maton A (1993) Human Biology and Health, Englewood Cliffs, New Jersey.

Patil M, Sheth KA, Adarsh CK (2013) Elevated alpha fetoprotein, no hepatocellular carcinoma. J Clin Exp Hepatol 3: 162-164.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS,

Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy number changes using cDNA microarrays. Nat Genet 23: 41-46.

Shen C and Liu ZP (2017) Identifying module biomarkers of hepatocellular carcinoma from gene expression data. Proceedings of 2017 Chinese Automation Congress (CAC), pp 5404-5407.

Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics. CA Cancer J Clin 64: 9-29.

Skawran B, Steinemann D, Weigmann A, Flemming P, Becker T, Flik J, Kreipe H, Schlegelberger B, Wilkens L (2008) Gene expression profiling of hepatocellular carcinoma: Upregulation of genes in amplified chromosome regions. Mod Pathol 21: 505-516.

Szala A, Sawicki S, Swierzko AS, Szemraj J, Sniadecki M, Michalski M, Kaluzynski A, Lukasiewicz J, Maciejewska A, Wydra D, *et al.* (2013) Ficolin-2 and ficolin-3 in women with malignant and benign ovarian tumours. Cancer Immunol Immunother 62: 1411-1419.

Timmermann A, Kuster A, Kurth I, Heinrich PC, Muller-Newen G (2002) A functional role of the membrane-proximal extracellular domains of the signal transducer gp130 in heterodimerization with the leukemia inhibitory factor receptor. Eur J Biochem 269: 2716-2726.

Tomida M (2000) Structural and functional studies on the Leukemia Inhibitory Factor Receptor (LIF-R): gene and soluble form of LIF-R, and cytoplasmic domain of LIF-R required for differentiation and growth arrest of myeloid leukemic cells. Leuk Lymphoma 37: 517-25.

Tung EK, Mak CK, Fatima S, Lo RC, Zhao H, Zhang C, Dai H, Poon RT, Yuen MF, Lai CL, *et al*. (2011) Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. Liver Int 31: 1494-1504.

Van de Steeg E, Stranecky V, Hartmannova H, Noskova L, Hrebicek M, Wagenaar E, Van Esch A, De Waart DR, Oude Elferink RP, Kenworthy KE, *et al*. (2012) Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. J Clin Invest 122: 519-528.

Wang W, Peng JX, Yang JQ, Yang LY (2009) Identification of gene expression profiling in hepatocellular carcinoma using cDNA microarrays. Dig Dis Sci 54: 2729-2735.

Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, *et al*. (2003) Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 9: 416-423.

Zakim D, Boyer TD (2003) Hepatology: A Textbook of Liver Disease, 4th ed., Saunders, Pennsylvania.