Data Mining on DNA Sequences of Hepatitis B Virus

Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie Y.T. Ng, Henry L.Y. Chan, Stephen K.W. Tsui, Tony S.K. Mok, Pete Chi-Hang Tse, and Joseph Jao-Yiu Sung

Abstract—Extraction of meaningful information from large experimental data sets is a key element in bioinformatics research. One of the challenges is to identify genomic markers in Hepatitis B Virus (HBV) that are associated with HCC (liver cancer) development by comparing the complete genomic sequences of HBV among patients with HCC and those without HCC. In this study, a data mining framework, which includes molecular evolution analysis, clustering, feature selection, classifier learning, and classification, is introduced. Our research group has collected HBV DNA sequences, either genotype B or C, from over 200 patients specifically for this project. In the molecular evolution analysis and clustering, three subgroups have been identified in genotype C and a clustering method has been developed to separate the subgroups. In the feature selection process, potential markers are selected based on Information Gain for further classifier learning. Then, meaningful rules are learned by our algorithm called the Rule Learning, which is based on Evolutionary Algorithm. Also, a new classification method by Nonlinear Integral has been developed. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral. The nonadditivity of the fuzzy measure reflects the importance of the feature attributes as well as their interactions. These two classifiers give explicit information on the importance of the individual mutated sites and their interactions toward the classification (potential causes of liver cancer in our case). A thorough comparison study of these two methods with existing methods is detailed. For genotype B, genotype C subgroups C1, C2, and C3, important mutation markers (sites) have been found, respectively. These two classification methods have been applied to classify never-seen-before examples for validation. The results show that the classification methods have more than 70 percent accuracy and 80 percent sensitivity for most data sets, which are considered high as an initial scanning method for liver cancer diagnosis.

Index Terms—Data mining, DNA sequences of HBV, mutation sites, nonlinear integrals, rule learning, the signed fuzzy measures.

1 Introduction

In Asia, infection of Hepatitis B virus (HBV) is a major health problem. At least 10 percent of the Chinese population (120 million people) are HBV carriers, and up to 25 percent of HBV carriers will die as a result of HBV-related complications including liver cirrhosis and hepatocellular carcinoma (HCC), i.e., liver cancer. Chronic infection by the HBV causes an increased risk of hepatocellular carcinoma (HCC) by more than 100-fold [1]. The relationship between HBV genotype and viral mutation with hepatocarcinogenesis is controversial. A case control study from Taiwan suggested that genotype C HBV is more closely associated with cirrhosis and HCC in those who are older than 50 years, whereas genotype B is more common in patients with HCC aged less than 50 years [2]. Our previous cohort study of 426 cases of chronic hepatitis B patients also reviewed a higher risk of HCC and liver cirrhosis in genotype C infection [3]. On the other hand, reports from Japan and China did not confirm the higher malignant potential of genotype C HBV [4], [5]. The aim of this study is to find the genomic markers of the HBV and clinical information which are useful in predicting occurrence of liver cancer and response to therapy.

In this study, we look into the clinical data prepared by the clinicians, and the HBV DNA genomes prepared by the biochemists of our research group [6], [7]. Patients taking part in this study were selected by the clinicians carefully, according to their age, sex, and past clinical status. Chronic hepatitis B patients recruited since 1997 were prospectively followed up for the development of HCC for avoiding selection bias. HCC was diagnosed by a combination of alpha fetoprotein, imaging, and histology. Liver cirrhosis was defined as ultrasonic features of cirrhosis together with hypersplenism, ascites, varices, and/or encephalopathy [3]. Clinical attributes for analysis were chosen by clinicians based on their expert knowledge. Primer Express software version 2.0 (PE applied Biosystems, Foster City, CA) was used to find suitable primers and probes. TaqMan real-time

- *K.S. Leung, K.H. Lee, and J.F. Wang are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. E-mail: {ksleung, khlee, jfwang}@cse.cuhk.edu.hk.*
- *E.Y.T. Ng's contact information is not available.*
- *H.L.Y. Chan is with the Department of Medicine and Therapeutics, The Chinese University of Hong Kong, 9/F Prince of Wales Hospital, Shatin, Hong Kong. E-mail: hlychan@cuhk.edu.hk.*
- *S.K.W. Tsui is with the Department of Biochemistry (Medicine), Institute of Science and Technology, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. E-mail: kwtsui@cuhk.edu.hk.*
- *T.S.K. Mok is with the Department of Clinical Oncology, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. E-mail: mok206551@cuhk.edu.hk.*
- *P.C.H. Tse is with the Department of Medicine and Therapeutics, Room 415, CC, PWH, Shatin, NT, Hong Kong. E-mail: petse@cuhk.edu.hk.*
- *J.J.Y. Sung is with the Department of Medicine and Therapeutics, Room 114009, 9/F, Clinical Sciences Building, Prince of Wales Hospital, Shatin, NT, Hong Kong. E-mail: joesung@cuhk.edu.hk.*

Manuscript received 23 June 2008; revised 4 Dec. 2008; accepted 7 Jan. 2009; published online 14 Jan. 2009.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-06-0116. Digital Object Identifier no. 10.1109/TCBB.2009.6.

PCR technology was used to differentiate the nucleotide variant [6]. Because the focus of this paper is on the study of data mining techniques, the selection process and criteria of patients and the research experiments run by our Biochemistry Department will not be discussed in detail.

In [8], HBV DNA sequences were taken from 13 patients. Keum et al. [9] amplified a conserved core region and a surface antigen region of HBV DNA by PCR from sera of 27 Korean chronic hepatitis B patients for detecting hepatitis B virus mutants. Our project is one of the biggest HBV DNA full-sequence collection and analysis studies of its kind. We have collected DNA sequences from 98 Control (normal) and 100 HCC (cancer) patients specifically for this project. The DNA sequences of HBV are not exactly the same for each group, and they possess some individual nucleotide mutations that may or may not be related to HCC. From previous studies, HBV can be divided into seven genotypes where each of them has more than 8 percent difference of nucleotides from the others. In Hong Kong, genotypes B and C are the predominant types, and all the examples we have are of these two genotypes. To reduce the noise of genotypic difference among the sequences collected, we propose to analyze these DNA examples in each genotype separately.

Classification is one of the most studied data mining tasks. The objective is to predict the value (the class) of a user-specified goal attribute based on the values of other attributes, called the predictive (feature) attributes. The goal attribute might be the prediction of whether or not a patient has cancer, while the predictive feature attributes might be the mutation sites of the patient's virus DNA.

The focus of this study is to identify genetic marker(s) for liver cancer (HCC) from HBV DNA sequences. There are similar medical research reports, but all of them are focused on the specific gene positions, proteins or part of a virus genome. However, our project is the first study on the complete viral genome. One of the past researches is an HIV genomic study [10]. The researchers align each DNA sequence with a reference sequence, and then select the genes using their expert knowledge, and use Decision Tree and Support Vector Machine (SVM) for analysis. In [11], the researchers focused on the identification of HBV DNA sequences that are predictive of response to one therapy. Some sites in sequences were observed to have caused the effect. Chan et al. studied the risk factors in HBV sequences with respect to medicine [3], [6]. Here, we apply soft computing tools to predict positive patients and analyze the effective mutation sites in the HBV DNA sequences.

The aim of this study is to develop a data mining framework which contains an appropriate classifier for liver cancer based on HBV DNA and clinical data. We develop two new algorithms based on rule learning (RL) and nonlinear Integral (NI). We then carry out a thorough comparative study on these two new models with existing classifiers. The classification model should have high sensitivity and acceptable accuracy and specificity for HCC diagnosis and prediction. The model learned should also give clear indication of the degrees of influence of the attributes toward the classification goal and whether there are any interactions among the predictive attributes. In this paper, we identified the important mutation sites (markers) in the HBV sequences that could have caused or been related to liver cancer. We use information entropy for finding genetic markers of HCC in the HBV genome data and propose a new classification model based on nonlinear integrals.

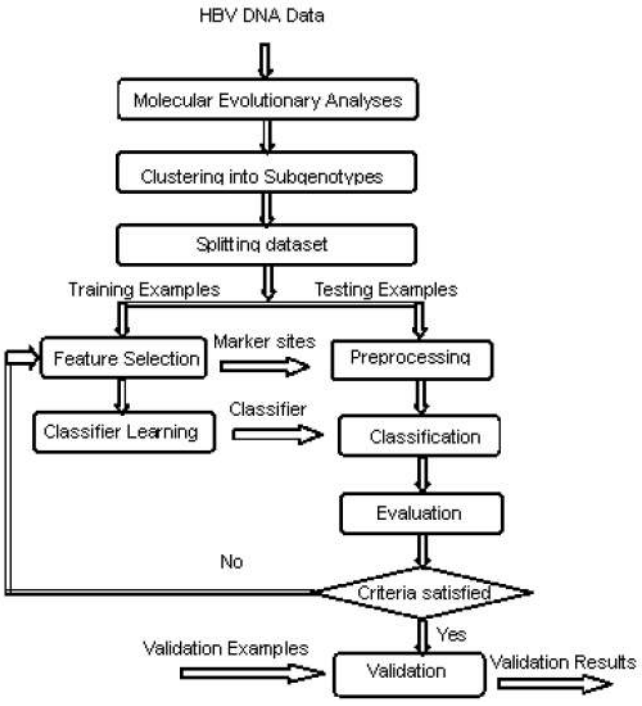

Fig. 1. Data mining framework.

This paper is organized as follows: Section 2 describes the data mining framework which includes the new rule learning and the nonlinear integral classification models in detail. All the methods and data sets used in this project are detailed in Section 3. The experimental results and the comparative studies are presented in Section 4. Section 5 concludes with the summary and the discussion of some directions for future work.

2 Data Mining Framework

The data mining framework developed is shown in Fig. 1. There are nine modules. After the molecular evolutionary analysis, the data are passed to the Clustering Module to check whether clusters exist based on the phylogenetic tree analysis. If clusters are found, each cluster will be analyzed separately for potential genetic marker sites because it will minimize the noise produced by the genotype differences and give much better classification accuracy. For each cluster (or genotype), the data are divided into training and test sets. The training examples are then passed to the Feature Selection Module to find the useful features (genetic marker sites) for classification. The potentially useful features (attributes) are extracted and passed to the Classifier Learning Module wherein a classifier is learned. The features selected are also sent to the preprocessing module to extract the values of these features in the testing data set for testing in the Classification module. Finally, the prediction results of the classifier are verified and evaluated based on the testing examples. If the evaluation results are unsatisfactory, i.e., stopping criteria are not satisfied, the learning process is

repeated starting from the feature selection; otherwise, the classifier will be validated by never-seen-before examples. The following sections will explain how the features are selected and also the basic principles of the classifier.

2.1 Molecular Evolutionary Analysis

Serum examples from 49 patients infected with HBV genotype C, as determined by previous genotype-specific restriction fragment length polymorphism analysis, were studied [3]. All serum examples were kept in a -80°C freezer for storage. All patients were ethnic Chinese and were followed up in the Hepatitis Clinic of the Prince of Wales Hospital (Hong Kong). All patients were positive for hepatitis B surface antigen for at least six months and had no evidence of hepatocellular carcinoma. Sixty-nine full-genome nucleotide sequences of HBV genotype C and 12 full-genome nucleotide sequences of nongenotype C HBV were also retrieved from the GenBank database for comparison. All reference sequences from GenBank were derived from patients with chronic hepatitis B; HBV nucleotide sequences from patients with acute hepatitis B hepatocellular carcinoma or patients treated with antiviral agents were excluded. The geographical origins of patients harboring different HBV genotype C genomes in GenBank were retrieved from the respective original publications and the descriptions in the GenBank database.

The full-genome nucleotide sequences of the isolates of HBV genotype C from our center were compared with those of the isolates of HBV genotype C and nongenotype C HBV retrieved from the GenBank database. Nucleotide sequences are multiple-aligned using ClustalW version 1.83 [12] and corrected manually by visual inspection. Genetic distances are estimated by Kimura's two-parameter method and the phylogenetic trees are constructed by the neighbor-joining method [13], [14]. The reliability of the pairwise comparison and phylogenetic tree analysis is assessed and assured by bootstrap resampling with 1,000 replicates. Phylogenetic and molecular evolutionary analyses are done using MEGA version 3.0 [15].

2.2 Clustering

Since different HBV subgroups are likely to be the result of divergence from genomic mutations over time and knowledge of the geographical distribution, genomic relatedness of the HBV genotype C subgroups will be useful in gaining an understanding of the spread of HBV in Asia. Hepatitis B virus genotype B (HBV/B) has been classified into five subgenotypes. In [16], a phylogenetic analysis of the complete genome sequences from the examples obtained from the Arctic and those from Japan and Asia revealed six distinct clusters within HBV/B. Within each HBV genotype C subgroup, several clusters with genomic resemblance to one another can be identified. The most well-defined example is the cluster in Okinawa, where the prevalence of HBV genotype C is much lower than that in the rest of Japan [17].

There are two genotypes, B and C, in the 200 plus HBV DNA sequences we collected specifically for this project. While genotype B HBV appears to be a homogenous group [18], the phylogenetic tree results show that there exist three main clusters in the genotype C among the HBV strains collected (Fig. 2) [7]. We label them as C1, C2, and C3,

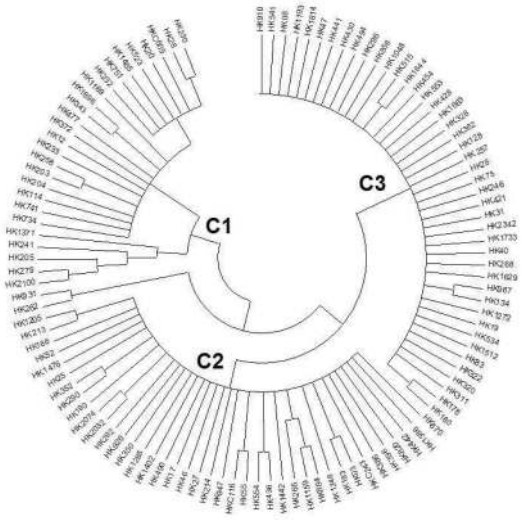

Fig. 2. Phylogenetic tree of genotype C.

respectively. Subgrouping of HBV genotype C was based on an intersubgroup difference of nucleotide sequence of > 4 percent [19]. This is in concordance with our previous phylogenetic analyses with published full-length sequence in the GenBank. The main reason for us to find markers separately from within the clusters (subgenotypes) obtained from clustering analysis is that these subgenotypes exhibit mutations (nucleotide site differences) caused by geographical diversity which are not markers for carcinogenic diagnosis. If we were to analyze all these subgenotype data as one genotype group, their intergenotypic differences would become distracting noises in the data mining process for markers. These three clusters can be identified by the combinations of four nucleotides. These three clusters will be analyzed separately in the classifier learning part.

2.3 Feature Selection Algorithm

The main purpose of feature selection [20], [21], [22], [23] is to reduce the number of features used in classification while maintaining acceptable classification accuracy. For example, the Sequential Forward Floating Selection (SFFS) algorithm proposed by Pudil et al. [24] was one of the commonly used algorithms [25]. The main advantage of this method is that it produces a hierarchy of feature subsets with the best selection for each dimension. However, we aim at global performance of the whole framework, so we adopt a simpler algorithm based on information gain to select initial features.

In our approach, information gain criterion [26] is used to find the useful features to distinguish between the Control (normal) and the HCC (cancer) groups of HBV carriers.

Information gain is a common criterion for feature selection. The information gain of a feature (attribute) is the uncertainty (entropy) that can be reduced if the attribute is used for classification. Hence, the information gain should be aimed at the higher the better. Equation (1) is the entropy E of an attribute X with m values, $X_1, X_2, \ldots, X_m$, and $P(X_i)$ is the probability of the value X_i

$$E(X) = \sum_{i=1}^{m} -P(X_i) log_2\ P(X_i). \quad (1)$$

Specific to a typical DNA classification problem, we assume that the data have K classes, $C = c_1, c_2, \ldots, c_K$. For each aligned site position, it has m possible nucleotides $V_1, V_2, \ldots, V_m$. We define $|c_k|, k = 1, 2, \ldots, K$, as the number of sequences in class c_k. $|c_{ki}|$ is the number of sequences in Class c_k whose character at the aligned site is V_i, which can be A, T, G, or C in our case. The Remainder of X, $R(X)$, is defined as follows:

$$R(X) = \sum_{i=0}^{m} \frac{\sum_{k=1}^{K} |c_{ki}|}{\sum_{k=1}^{K} |c_k|} E(P(c_{1i}), \ldots, P(c_{ki})). \quad (2)$$

Information Gain IG_j of the aligned site j is the difference between the original information content $E(C)$ of the data set and the amount of information needed to classify all the unclassified data left in the data set after applying site j for classification

$$IG_j = E(C) - R(j). \quad (3)$$

The features are ranked by the information gains, and then the top-ranked features are chosen as the potential attributes used in the classifier. A site with higher information gain will contribute more in the classification and be able to distinguish more examples (cases).

2.4 Current Classification Algorithms

There are several common classification models such as Naïve Bayesian Network [27], [28], [29], Decision Tree, Neural Networks, and Rule Learning using Evolutionary Algorithm [30]. The learning processes of Naïve Bayesian Networks and Decision Tree are faster. However, they cannot cope well with feature interactions. Neural Networks are treated as black box learning and it is difficult for a human to understand or interpret the classification explicitly.

However, Rule Learning using Evolutionary Algorithm performs a global search and can cope with feature interactions better than the previous classification models [31], [32]. Also, the classification rules generated are simple and easily interpretable by human experts who frequently use the same reasoning approach very much similar to the rules. Therefore, the Rule Learning Using Evolutionary Algorithm approach is clearly a better choice in terms of interpretability of the knowledge acquired through the classifier learned.

Rule learning tries to learn rules from a set of training data (examples). It can be modeled as a search problem of finding the best rules that classify the training examples with minimum error. However, the search space can be very large; a robust search algorithm is required. Here, Generic Genetic Programming (GGP) [33], [34], which is a type of the Evolutionary Algorithms (EA), is adopted as our search and optimization algorithm. First, a population is initialized by generating individuals (a set of rules) randomly. A fitness function is used to evaluate how good an individual is, that is, how many cases it can classify correctly. Then, some individuals are selected to evolve (generate) new individuals with the genetic operators. Individuals become better and better through the evolution process until the termination criterion is met.

The input is the training data set, and the output is a rule set, which can classify the training data with higher accuracy. We assume that there are n features (attributes), $X = x_1, x_2, \ldots, x_n$, and K classes, $C = c_1, c_2, \ldots, c_K$. For each attribute x_j, one of its m_j values can be taken. Each rule includes two components: the antecedent (IF part) and the consequence (THEN part), as follows:

IF $(x_1 = v_1) \wedge (x_2 = v_2) \wedge \cdots \wedge (x_l = v_l)$ THEN Class is $C = c_k$, where the antecedent includes l $(l = [1, n])$ unique attributes $x_1, x_2, \ldots, x_l \in \{x_1, x_2, \ldots, x_n\}$, $v_1, v_2, \ldots, v_l \in \{A, G, T, C\}$, and $c_k, k \in \{1, 2, \ldots, K\}$, is a certain class to which the object is to be classified. In our case, we have only two classes, namely HCC and CONTROL. There are l unique attributes present in and $n - l$ attributes absent from each rule. Each attribute present in the antecedent can only take one of its possible values, $\{A, G, T, C\}$. All the rules in the output rule set are connected by ELSE IF, meaning that the order of application of the rules must be followed.

We use a simple example to illustrate the rules deduced by the Rule Learning. For HBV data set B, which will be introduced in the following section, we have learned the rules for diagnosing liver cancer (HCC) and nonliver cancer (CONTROL) cases. The rules are given as follows:

IF A1762 and G1764 and C53, then HCC,
ELSE IF T1762 and A1764 and CG2712, then HCC,
ELSE IF T1762 and A1764 and T2712 and C2525, then HCC,
ELSE CONTROL.

Although Rule learning based on EA can interpret the interaction of features, the degree of the interaction cannot be analyzed exactly by a measure. Hence, we introduce the Fuzzy Measure to describe the interaction with respect to the classification. A new classification model is proposed based on Nonlinear Integrals with respect to signed Fuzzy Measure in the following section.

2.5 Classification Based on Nonlinear Integrals

In classification, we are given a data set consisting of N example records, called the training set, where each record contains the value of a classifying attribute Y and the value of feature attributes $x_1, x_2, \ldots, x_m$. Positive integer N is the data size. The classifying attribute indicates the class to which each example belongs, and it is a categorical attribute with values coming from an unordered finite domain. The set of all possible values of the decisive attribute is denoted by $C = c_1, c_2, \ldots, c_K$, where each c_k, $k = 1, 2, \ldots, K$, refers to a specified class. The feature attributes are numerical, and their values are described by an m-dimensional vector, $(f(x_1), f(x_2), \ldots, f(x_m))$. The range of the vector, a subset of n-dimensional euclidean space, is called the feature space. Thus, the j example record consists of the jth observation for all feature attributes and the classifying attribute, and is denoted by $(f_j(x_1), f_j(x_2), \ldots, f_j(x_m), Y_j), j = 1, 2, \ldots, N$ [35].

In this section, a method of classification based on nonlinear integrals will be presented. It can be viewed as an idea of projecting the points in the feature space onto a real axis through a nonlinear integral, and then using a one-dimensional classifier to classify these points according to a certain criterion optimally. Our classifying attributes holding the discrete value of A, C, G, or T is numericalized to be a virtual variable. All of these are realized under the guidance of an adaptive genetic algorithm [36]. Good

performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral, since the nonadditivity of the fuzzy measure reflects the importance of the feature attributes, as well as their inherent interactions, toward the discrimination of the points. In fact, each feature attribute has its, respective, important index reflecting its amount of contribution toward the decision. Furthermore, the global contribution of several feature attributes to classification is not just the simple sum of the contribution of each feature to the decision, but may vary nonlinearly. A combination of the feature attributes may have a mutually restraining or a complementary synergy effect on their contributions toward the classification decision. Hence, the fuzzy measure defined on the power set of all feature attributes is a proper representation of the respective importance of the feature attributes and the interactions among them, and a relevant nonlinear integral is a good fusion tool to aggregate the information coming from the individual and the combinations of the feature attributes for the classification. The following are the details of these basic concepts and the mathematical model for the classification problem.

2.5.1 Fuzzy Measures and Nonlinear Integrals

Let $X = x_1, x_2, \ldots, x_m$ be a nonempty finite set of feature attributes and $P(X)$ be the power set of X.

Definition 3.1. *A fuzzy measure μ is a mapping from $P(X)$ to $[0, \infty)$ satisfying the following conditions:*

1. $\mu(\phi) = 0$ *and*
2. $A \subset B \Rightarrow \mu(A) \leq \mu(B), \forall A, B \in P(X)$.

The set function μ is nonadditive in general. If $\mu(X) = 1$, then μ is said to be regular.

To further understand the practical meaning of the fuzzy measure, let us consider the elements in a universal set X as a set of predictive attributes to predict a certain objective. Then, for each individual predictive attribute as well as each possible combination of the predictive attributes, a distinct value of a fuzzy measure is assigned to describe its influence to the objective. Due to the nonadditivity of the fuzzy measure, the influences of the predictive attributes on the objective are dependent in a manner such that the global contribution of them to the objective is not just the simple sum of their individual contributions.

Example 3.1. Assume that we have observed three symptoms of a patient and want to determine which disease he or she is suffering from. The symptoms are regarded as the information sources, which form the universal set denoted by $X = \{x_1, x_2, x_3\}$. Their individuals as well as joint influences on the prediction of disease are specified by a fuzzy measure μ defined in Table 1.

TABLE 1
An Example of Fuzzy Measure Defined on $X = \{x_1, x_2, x_3\}$

Set	Value of μ	Set	Value of μ
Ø	0.0	$\{x_3\}$	0.2
$\{x_1\}$	0.2	$\{x_1, x_3\}$	0.5
$\{x_2\}$	0.4	$\{x_2, x_3\}$	0.9
$\{x_1, x_2\}$	0.5	$\{x_1, x_2, x_3\}$	1.0

Here, $\mu(\{x_2, x_3\}) > \mu(\{x_2\}) + \mu(\{x_3\})$ indicates that the joint contribution of x_2 and x_3 to the diagnosis is greater than the sum of their individual contributions. This shows that the interaction between x_2 and x_3 enhances the influence of each other. On the other hand, $\mu(\{x_1, x_2\}) < \mu(\{x_1\}) + \mu(\{x_2\})$ shows that x_1 and x_2 are restraining each other. Note that the essential properties of fuzzy measure are monotonicity and vanishing at the empty set. This implies that fuzzy measure only allows its value to be nonnegative.

However, the monotonicity and nonnegativity of fuzzy measure are too restrictive for real applications. In this paper, we assume that μ is a signed fuzzy measure on $P(X)$, i.e., $\mu : P(X) \rightarrow (-\infty, +\infty)$ and $\mu(\phi) = 0$. For convenience, $\mu(\{x_1\}), \mu(\{x_2\}), \ldots, \mu(\{x_n\}), \mu(\{x_1, x_2\}), \ldots, \mu(\{x_1, x_2, \ldots, x_n\})$ are sometimes abbreviated as $\mu_1, \mu_2, \ldots \mu_n, \mu_{12}, \ldots, \mu_{12\cdots n}$, respectively.

Definition 3.2. *Let f be a nonnegative function on X. The Nonlinear integral of f with respect to μ, $\int f d\mu$, is defined by*

$$\int f d\mu = \int_0^\infty \mu(F_\alpha) d\alpha,$$

where $F_\alpha = \{x | f(x) \geq \alpha\}$ for any $\alpha \in [0, \infty)$, is the α-level set of f.

To calculate the value of the Nonlinear integral of a given function f, the values of f, $\{f(x_1), f(x_2), \ldots, f(x_m)\}$, should be first arranged into an increasing order, that is,

$$f(x'_1) \leq f(x'_2) \leq \cdots \leq f(x'_m),$$

where $(x'_1, x'_2, \ldots x'_m)$ is a certain permutation of $(x_1, x_2, \ldots x_m)$. Then, the value of the Nonlinear Integral can be obtained by computing

$$(c) \int f d\mu = \sum_{i=1}^{m} [f(x'_i) - f(x'_{i-1})] \mu(\{x'_i, x'_{i+1}, \ldots x'_m\}),$$

where $f(x'_0) = 0$.

2.5.2 Nonlinear Integral Projection Based on the Nonlinear Integral

We can build an aggregation tool that projects the feature space onto a real axis. Under the projection, each point in the feature space becomes a value of the virtual variable.

A point $(f(x_1), f(x_2), \ldots, f(x_m))$, denoted by $(f_1, f_2, \ldots, f_m)$, simply, in the feature space can be regarded as a function f defined on X. It represents an observation of the feature attributes. Point f is projected to be $\hat{Y}$, the value of the virtual variable, on a real axis through a nonlinear integral defined by

$$\hat{Y} = \int f d\mu.$$

Once the value of μ is determined, we can calculate the virtual value $\hat{Y}$ from f.

2.5.3 GA-Based Adaptive Classifier

The next step is to find an appropriate formula that projects the n-dimensional feature space onto a real axis L such that each point $f = (f_1, f_2, \ldots, f_m)$ becomes a value of the virtual variable that is optimal with respect to the classification. In such a way, each classification boundary is just a point on real axis L.

The classification process can be divided into two parts for implementation:

Step 1. The nonlinear integral classifier depends on the fuzzy measure μ, so the first step is to determine the optimal values of μ by using the GA tool. In fact, the fitness function comes from the linear classifier used in the second procedure. It is an iterative process. The optimal Fuzzy Measure will be the output to the next step.

Step 2. When the fuzzy measure μ is determined, the virtual value can be obtained using the Nonlinear Integral. Then, we can classify these virtual values on real axis using a linear classifier.

The following section focuses on the above problems.

GA-based learning fuzzy measure. Here, we discuss the optimization of the fuzzy measure μ under the criterion of minimizing the corresponding global misclassification rate, which is obtained in the second part above.

In our GA model, we use a variant of the original function f, $f' = a + bf$, where a is a vector to shift the coordinates of the data and b is a vector to scale the values of predictive attributes. Each chromosome represents fuzzy measure vector μ, shifting vector a, and scaling vector b. A signed fuzzy measure is 0 at empty set. If there are m attributes in training data, a chromosome has $2^m - 1 + 2m$ genes which are set to random real values at initialization. Genetic operations used are traditional ones. At each generation, for each chromosome, all variables are fixed and the virtual values of all training data are calculated using a nonlinear integral. The fitness function can be defined as follows:

$$fitness = \omega_1 * accuracy + \omega_2 * sensitivity,$$

where ω_1 and ω_2 are the adjustment parameters given by users. Accuracy and sensitivity are determined in the second part of the model.

Linear classifier for the virtual values. After determining the fuzzy measure μ, shifting vector a, scaling vector b, and the respective classification function from the training data in GA, points in the m-dimensional feature space are projected onto a real axis using a nonlinear integral.

We use Fisher's linear discriminant function to perform classification in this one-dimensional space [37], [38]. Positive and negative centroids for projected data are determined by the following formulas:

$$m_+ = \frac{\sum_{i:y_i=1} x_i}{\sum_{i:y_i=1} 1}, \qquad m_- = \frac{\sum_{i:y_i=-1} x_i}{\sum_{i:y_i=-1} 1}.$$

Ronald Fisher defined the scatter matrices as

$$S_\pm \equiv \sum_{x_i:y_i=\pm 1} (x_i - m_\pm)(x_i - m_\pm)'.$$

$S_W = S_+ + S_-$ is called the Within-Class Scatter Matrix. Similarly, the Between-Class Scatter Matrix can be defined as

$$S_B \equiv (m_+ - m_-)(m_+ - m_-)'.$$

Hence, this result in an equivalent expression for Fisher's discriminant criterion is a ratio between two quadratic forms as

$$J(w) = \frac{w' S_B w}{w' S_W w},$$

in which w represents the direction of the projection space, i.e., the one-dimensional space. We can solve the programming problem by maximizing $J(w)$. The optimal w can be represented as $w = S_w^1 * (m_+ - m_-)$. So, the Fisher's discriminant function is formulated as

$$y = w * (x - n_+ * m_+ - n_- * m_-),$$

in which $n_\pm$ is the sum of observations in each class, respectively. Finally, a threshold needs to be fixed in order to define a complete classifier.

3 Methods

We applied EA-based Rule Learning [39] and Nonlinear Integral classifiers to classify the HBV DNA data into liver cancer (HCC) and normal (CON, control) classes, and then compare them with several classical classification methods which include See5.0 (Decision Tree) [40], Neural Network [41], SVM [42], and Naïve Bayes [43]. As mentioned before, we conduct a detailed study on the Rule Learning and Nonlinear Integral classifier separately. In this section, we will give brief descriptions about these classical methods of classification and the data sets used. Then, the implementation details of the Nonlinear Integral (NI) classifier and the evaluation methodology will be introduced.

3.1 Methods Description

The following paragraphs are the brief descriptions of the four classical methods used to compare with our new methods.

3.1.1 Decision Tree [40]

A decision tree is a tree-structured classifier. The Decision Tree method learns decision tree using a recursive tree growing process. Each test corresponding to an attribute is evaluated on the training data using a test criteria function. The test criteria function assigns each test a score based on how well it partitions the data set. The test with the highest score is selected and placed at the root of the tree. The subtrees of each node are then grown recursively by applying the same algorithm to the examples in each leaf. The algorithm terminates when the current node contains either all positive or all negative examples. We used the widely available package—See5.0, which is the state-of-the-art of the Decision Tree classifier.

3.1.2 Neural Network [41]

An Artificial Neural Network (ANN), or commonly just called neural network (NN), is an interconnected group of

TABLE 2
GATool Parameters in Matlab

Parameter	Set value	Parameter	Set value
PopulationType	doubleVector	Stall generations	Inf
PopulationSize	20	Stall time limit	Inf
EliteCount	2	Tolerance	1.0000e-006
CrossoverFraction	0.8000	Constraint tolerance	1.0000e-006
CrossoverFunction	@crossoverscattered	InitialPenalty	10
MigrationDirection	'forward'	PenaltyFactor	100
MigrationInterval	20	PlotInterval	1
MigrationFraction	0.2000	CreationFunction	@gacreationuniform
Generations	100	FitnessScalingFunction	@fitscalingrank

artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. In most cases, an NN is an adaptive system that changes its structure or weights of the interconnections based on external and internal information (stimuli) that flows through the network.

In more practical terms, NNs are nonlinear statistical data modeling for decision-making and classification tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. However, it is essentially a black box approach, and it is not easy to interpret how they function.

3.1.3 Support Vector Machine [42]

SVMs are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since, in general, the larger the margin the smaller the generalization error of the classifier.

The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. The classification model produced by SVM (as described above) only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. In our paper, the software used is obtained from [42].

3.1.4 Naïve Bayes [43]

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model."

Depending on the precise nature of the probability model, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the naïve Bayes model without believing in Bayesian probability or using any Bayesian method.

In spite of their oversimplified assumptions, Naïve Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficiency of Naïve Bayes classifiers [44]. An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

TABLE 3
The Details of HBV Data Sets

Dataset	Control	HCC	Total	%
B	51	37	88	43.878
C1	10	16	26	13.265
C2	18	22	40	20.408
C3	19	25	44	22.449
Total	98	100	198	

3.2 Implementation Details of Nonlinear Integral

To implement the learning algorithm of our new classifier based on nonlinear integrals, we use the GA tool in Matlab v7.2 Programming combined with Fisher's discriminant function programming [45]. All the parameters of our GA in our experiments are shown in Table 2. We set the generation limit to be 100 as the stopping criteria.

3.3 Data Description

The data set contains 98 control patients and 100 HCC patients. The HBV DNA sequences are obtained specifically for this study from these patients carefully selected by our medical experts to minimize the demographic bias. There are four data sets corresponding to the different clusters, namely B, C1, C2, and C3. The numbers of patients for each data set are shown in Table 3 in which the last column represents the proportion of each data set. For each data set, an independent validation set is prepared to evaluate the performance of the classifiers. Table 4 shows the number of patients of the validation data sets.

TABLE 4
Summary of Validation Sets

Datasets	Control	HCC	Total
B	8	7	15
C1	7	5	12
C2	9	6	15
C3	5	5	10
Total	29	23	52

3.4 Evaluation Methodology

In classifying an unknown case, depending on the class predicted by the classifier and the true class of the patient (Control or HCC), four possible types of results can be observed for the prediction as follows:

1. True positive—the result of the patient has been predicted as positive (Cancer) and the patient has cancer.
2. False positive—the result of the patient has been predicted as positive (Cancer) but the patient does not have cancer.
3. True negative—the result of the patient has been predicted as negative (Control), and indeed, the patient does not have cancer.
4. False negative—the result of the patient has been predicted as negative (Control) but the patient has cancer.

Let TP, FP, TN, and FN, respectively, denote the number of true positives, false positives, true negatives, and false negatives. For each learning and evaluation experiment, *Accuracy*, *Sensitivity*, and *Specificity* defined below are used as the fitness or performance indicators of the classification:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN),$$
$$Sensitivity = TP/(TP + FN),$$
$$Specificity = TN/(TN + FP).$$

For screening tests, medical professionals usually will prefer to have higher sensitivity, i.e., lower accuracy and specificity is an acceptable trade-off for high sensitivity as long as the accuracy and specificity are reasonable. It means that we rather send more people for confirmation tests than miss any true cancer patients. In the data sets, all attributes are categorical attributes. There are four symbolic values A, C, G, and T for each attribute. In order to use the nonlinear model, we use simple integer values, 0, 1, 2, and 3, as the numericalized initial values to represent the discrete values of the attributes, respectively.

We adopt K-fold cross-validation method to make sure that the whole data set can be used as testing data in turn and overtraining (overfitting) can be avoided. It means that we randomly partition the n data into K sets with size of n/K, which are trained on $(K-1)$ sets and tested on the remaining set, and repeated K times in turn thus taking the mean result. After K runs, all data are used for testing and the average can be computed to evaluate the performance. The K-fold method is repeated 10 times for each experiment ($10 \times K$ runs in total) to obtain an overall average performance.

Our data sets are very small despite the fact that they are one of the biggest single studies. For example, C1 contains DNA sequences from 26 individuals. We must ensure that there is at least one positive case for each class in the testing data set. If the number (K) of splits is too large, the size of the testing set will be too small, and it may not even have a positive case. On the other hand, if the numbers of splits are too small, it will result in small training sets which may not contain sufficient information for training. So, we need to find a balance between the sizes of training and testing sets in order to reduce the probability of overtraining (overfitting) and undertesting (i.e., not enough positive and negative examples for testing). We have tried several feasible K values for Nonlinear Integrals. The results are shown in Table 5. We can see that the testing accuracy and sensitivity are best by taking 10-fold. Consequently, we have chosen a 10-fold method for our experiments. This 10-fold methodology is applied to all experiments including the classical classifier used in our comparison.

TABLE 5
All Splits for Training and Testing on Nonlinear Integral Classifier

Data		2fold	3fold	4fold	5fold	6fold	7fold	8fold	9fold	10fold
B training	Acc	0.747	0.731	0.728	0.728	0.730	0.722	0.723	0.725	0.682
	Sen	0.808	0.829	0.805	0.803	0.804	0.814	0.802	0.802	0.811
	Spe	0.702	0.660	0.672	0.672	0.677	0.656	0.665	0.669	0.588
B testing	Acc	0.649	0.646	0.647	0.643	0.621	0.637	0.634	0.625	0.674
	Sen	0.687	0.700	0.687	0.678	0.651	0.698	0.696	0.678	**0.813**
	Spe	0.621	0.606	0.617	0.615	0.600	0.593	0.588	0.587	0.574
C1 training	Acc	0.962	0.961	0.962	0.961	0.960	0.962	0.962	0.962	0.961
	Sen	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
	Spe	0.900	0.899	0.900	0.900	0.896	0.900	0.900	0.900	0.899
C1 testing	Acc	0.785	0.815	0.840	0.843	0.848	0.845	0.856	0.849	0.847
	Sen	0.963	0.952	0.944	0.963	0.964	0.957	0.975	0.972	**0.980**
	Spe	0.500	0.594	0.683	0.660	0.650	0.650	0.656	0.661	0.640
C2 training	Acc	0.905	0.893	0.883	0.876	0.877	0.875	0.876	0.870	0.918
	Sen	0.895	0.883	0.882	0.873	0.878	0.884	0.873	0.872	0.903
	Spe	0.915	0.903	0.883	0.879	0.876	0.866	0.878	0.869	0.937
C2 testing	Acc	0.725	0.721	0.730	0.723	0.723	0.715	0.730	0.728	0.817
	Sen	0.645	0.624	0.655	0.650	0.660	0.669	0.648	0.657	0.788
	Spe	0.805	0.818	0.805	0.795	0.786	0.762	0.813	0.798	0.860
C3 training	Acc	0.780	0.774	0.772	0.767	0.765	0.761	0.765	0.761	0.731
	Sen	0.724	0.753	0.738	0.715	0.743	0.739	0.717	0.738	0.913
	Spe	0.853	0.803	0.816	0.835	0.795	0.792	0.828	0.794	0.491
C3 testing	Acc	0.651	0.609	0.606	0.624	0.604	0.627	0.624	0.606	0.600
	Sen	0.565	0.529	0.560	0.564	0.552	0.575	0.570	0.546	0.738
	Spe	0.763	0.712	0.670	0.698	0.674	0.695	0.698	0.670	0.410
Weighted training AVE.	Acc	0.814	0.803	0.799	0.797	0.798	0.793	0.794	0.793	0.777
	Sen	0.832	0.845	0.831	0.824	0.831	0.836	0.824	0.828	**0.877**
	Spe	0.805	0.772	0.777	0.780	0.772	0.761	0.775	0.767	0.678
Weighted testing AVE.	Acc	0.683	0.675	0.680	0.681	0.668	0.678	0.681	0.671	**0.709**
	Sen	0.688	0.680	0.686	0.685	0.672	0.699	0.695	0.683	**0.813**
	Spe	0.674	0.671	0.676	0.676	0.660	0.657	0.667	0.658	0.604

Note: Acc=Accuracy; Sen=Sensitivity; Spe=Specificity

4 Experimental Results

In this section, we first present the results of EA-based Rule Learning [39] and Nonlinear Integral classifiers to classify the HBV DNA data into liver cancer (HCC) and normal (CON, control) classes, and then compare them with several traditional classification methods which include See5.0 (Decision Tree) [40], Neural Network [41], SVM [42], and Naïve Bayes [43]. As mentioned before, we do a detailed study on the Rule Learning and Nonlinear Integral classifier separately because of the importance of their high interpretability of the models representing the knowledge acquired through the learning processes. The biochemists and doctors can see explicitly and clearly the influences of the mutated sites or markers and their potential interactions toward the formation of liver cancer.

For each data set, we will use the five attributes which include those selected by the Rule Learning method, and in some cases, supplemented by those with the highest

TABLE 6
Comparison Results of Nonlinear Integral with and without Feature Selection

Datasets \ Performance		With FS		Without FS	
		Train	Test	Train	Test
B	Accuracy	0.771	0.683	0.687	0.617
	Sensitivity	0.858	0.738	0.337	0.233
	Specificity	0.708	0.646	0.942	0.893
C1	Accuracy	0.922	0.790	0.826	0.515
	Sensitivity	1.000	0.920	0.937	0.715
	Specificity	0.798	0.600	0.650	0.190
C2	Accuracy	0.871	0.750	0.793	0.626
	Sensitivity	0.888	0.700	0.813	0.698
	Specificity	0.854	0.800	0.769	0.530
C3	Accuracy	0.828	0.732	0.813	0.750
	Sensitivity	0.843	0.705	0.718	0.643
	Specificity	0.808	0.765	0.937	0.900

information gain obtained by Viewer [46] partially shown in Fig. 9. For reducing computational complexity, we reduce the number of attributes by including the feature selection method. We compared the results of Nonlinear Integral with and without feature selection in Table 6. It shows that feature selection is very useful.

4.1 Comparison between NIC and RL

Table 7 shows the comparison results of Rule Learning and the NIC, and Table 8 shows the comparison results of our methods with several classic methods on data sets B, C1, C2, and C3. The results of RL and NIC for each data set and a validation set, which contains the never-seen-before cases, are shown in Table 7.

In Table 7, sensitivity results of NIC are higher than those of RL in most cases and other values are comparable. Since sensitivity is more important for doctors to diagnose, the performance of NIC is considered to be better than that of RL. Furthermore, NIC can not only determine the important sites (markers) with regard to the diagnosis but also give their degrees of contribution in real values, which are relatively meaningful in biomedical research. This will be described in the following section.

TABLE 7
Results of RL and NIC for Each Data Set

Datasets \ Performance		RL			NIC		
		Train	Test	Valid -ation	Train	Test	Valid -ation
B	Accuracy	0.716	**0.716**	0.769	0.771	0.683	0.721
	Sensitivity	0.730	0.731	0.800	0.858	**0.738**	0.742
	Specificity	0.706	0.707	0.750	0.708	0.646	0.697
C1	Accuracy	0.808	**0.800**	0.917	0.922	0.790	0.712
	Sensitivity	0.812	0.790	1.000	1.000	**0.920**	0.854
	Specificity	0.800	0.800	0.857	0.798	0.600	0.570
C2	Accuracy	0.775	**0.775**	0.917	0.871	0.750	0.712
	Sensitivity	0.700	**0.700**	1.000	0.888	**0.700**	0.854
	Specificity	0.850	0.850	0.857	0.854	0.800	0.570
C3	Accuracy	0.773	**0.770**	0.647	0.828	0.732	0.639
	Sensitivity	0.720	**0.717**	0.700	0.843	0.705	0.721
	Specificity	0.842	0.835	0.571	0.808	0.765	0.523

Note: RL=Rule Learning; NIC=Nonlinear Integral Classifier

TABLE 8
Comparison Results with Classical Methods for All Data Sets

Datasets \ Algorithms		NN	DT	NB	SVM	**NIC**	RL
B training	Accuracy	0.681	0.682	0.689	0.674	0.682	0.716
	Sensitivity	0.805	0.811	0.790	0.794	0.811	0.730
	Specificity	0.591	0.588	0.617	0.589	0.588	0.706
B testing	Accuracy	0.680	0.681	0.650	0.680	0.674	0.716
	Sensitivity	0.806	0.812	0.758	0.795	0.813	0.731
	Specificity	0.589	0.571	0.573	0.597	0.574	0.707
C1 training	Accuracy	0.889	0.937	0.894	0.897	0.961	0.808
	Sensitivity	1.000	1.000	0.722	0.965	1.000	0.812
	Specificity	0.711	0.836	1.000	0.810	0.899	0.800
C1 testing	Accuracy	0.869	0.717	0.650	0.961	0.847	0.800
	Sensitivity	0.999	1.000	0.300	1.000	0.980	0.790
	Specificity	0.677	0.280	0.850	0.899	0.640	0.800
C2 training	Accuracy	0.805	0.839	0.773	0.725	0.918	0.775
	Sensitivity	0.799	0.749	0.993	0.665	0.903	0.700
	Specificity	0.813	0.953	0.589	0.785	0.937	0.850
C2 testing	Accuracy	0.746	0.728	0.727	0.848	0.817	0.775
	Sensitivity	0.715	0.615	0.897	0.789	0.788	0.700
	Specificity	0.783	0.880	0.592	0.907	0.860	0.850
C3 training	Accuracy	0.628	0.684	0.697	0.604	0.731	0.773
	Sensitivity	0.707	0.504	0.688	0.475	0.913	0.720
	Specificity	0.524	0.905	0.702	0.780	0.491	0.842
C3 testing	Accuracy	0.573	0.645	0.587	0.753	0.600	0.770
	Sensitivity	0.619	0.442	0.600	0.663	0.738	0.717
	Specificity	0.524	0.920	0.567	0.871	0.410	0.835
Weight Average training	Accuracy	0.722	0.748	0.735	0.698	**0.777**	0.753
	Sensitivity	0.808	0.755	0.799	0.720	**0.877**	0.732
	Specificity	0.637	**0.765**	0.680	0.700	0.678	0.778
Weight Average testing	Accuracy	0.695	0.687	0.652	**0.767**	0.709	0.751
	Sensitivity	0.772	0.715	0.691	0.791	**0.813**	0.729
	Specificity	0.625	0.673	0.612	0.760	0.604	**0.777**

Note: NN=Neural Network; DT=Decision Tree; NB=Naïve Network; SVM=Support Vector Machine; RL=Rule Learning; NIC=Nonlinear Integral Classifier.

4.2 Results of Classifier Based on Nonlinear Integrals Compared with Other Methods

Table 8 shows the comparison results of the Nonlinear Integrals Classifier with five classical algorithms which include NN, Decision Tree (DT), Naïve Network (NB), SVM, and RL.

We run six sets of experiments for each classifier. The first set of experiments uses the top one site (attribute with the highest information gain), the second set the top two sites, the third one uses top three sites, and so on. The results are evaluated mainly according to the test accuracy and sensitivity for HBV data. Finally, the best result out of the six sets of experiments for each method is selected for comparison. Thus, we can see the optimal results of the NIC method based on GA compared against the other methods in Table 8.

For weighted average results in Table 8, the best ones are bolded. The weighted average is computed according to the number of cases in each data set. The sensitivity and accuracy of our NIC is better than most algorithms or is at least comparable. For comparing the performance of all algorithms graphically, we plot all the results in Figs. 3, 4, 5, 6, and 7. Meanwhile, we place all methods on an ROC space as in Fig. 8 for helping interpret the results in Table 8.

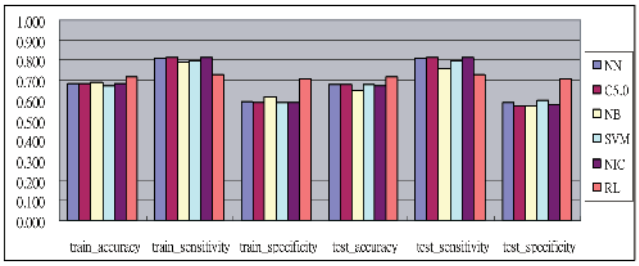

Fig. 3. The comparison of all methods for B.

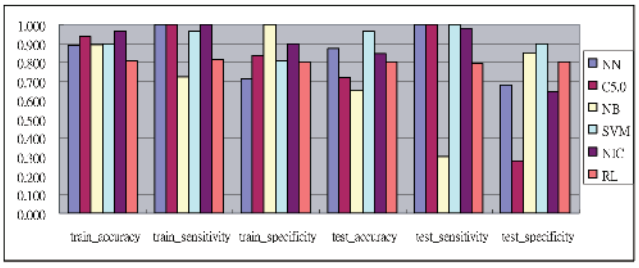

Fig. 4. The comparison of all methods for C1.

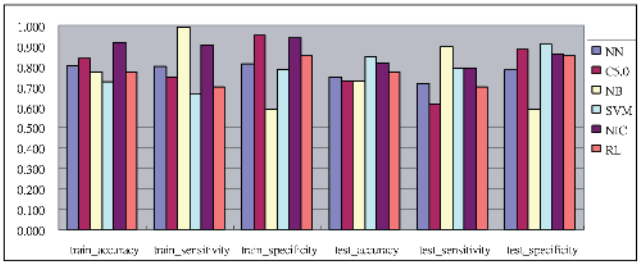

Fig. 5. The comparison of all methods for C2.

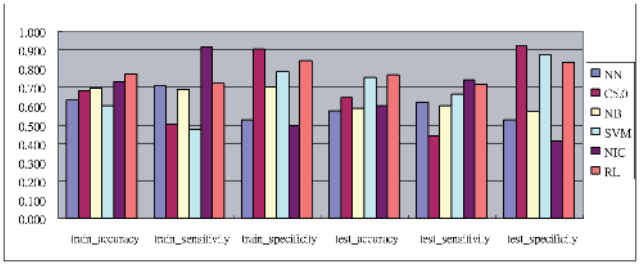

Fig. 6. The comparison of all methods for C3.

4.3 Comments on Results

Our framework includes RL and NI. RL has slightly higher accuracy than NI. It means that this method can have higher prediction power. But for doctors and clinicians, the sensitivity is more important than the accuracy, and NI is better than RL on sensitivity.

Compared with the four traditional methods, namely, NN, DT, NB, and SVM, NI shows the best diagnostic performance on the average evaluations. NI not only has comparable accuracy, it can also show the interaction of attributes. How to identify the importance of attributes and their combinations will be introduced in the next section.

4.4 To Identify Important Sites and Interactions among Them

Another important contribution of the nonlinear integral classifier is that we can find some significant sites (markers) and interactions among them in the sequences for further wet laboratory analyses. According to the definition of Nonlinear Integrals, for each data set, we can get a set of linear equations about the signed fuzzy measures as variables. A solution with the fewest nonzero values can be obtain by solving linear equations based on L1-Norm regularization [47].

The respective potential sites according to information gain in the sequences for x_i computed by Viewer [46] are listed in Table 9. Fig. 9 is the screenshot from the Viewer for Data Set C1. The left column represents the site numbers and the right column is the corresponding information gain values ranked in decreasing order.

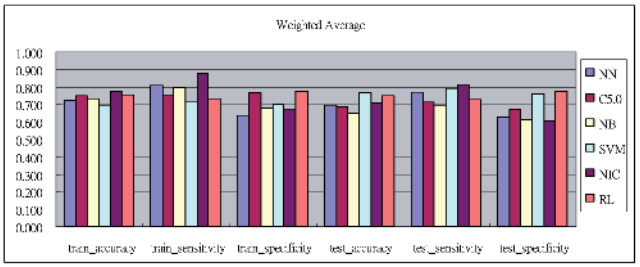

Fig. 7. The comparison of all methods as weighted average.

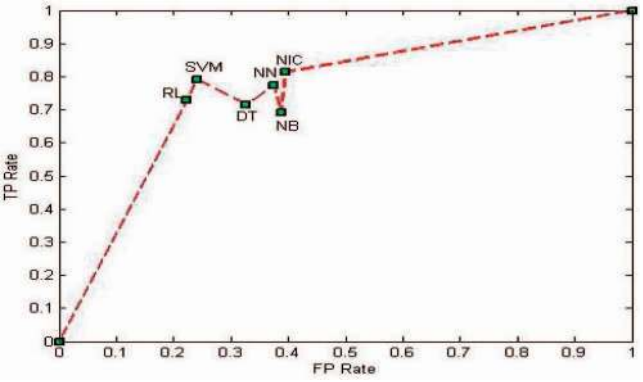

Fig. 8. Classifiers in ROC space.

TABLE 9
The Top 5 Sites No. of Sequences for Each Data Set

Data	x_1	x_2	x_3	x_4	x_5
B	1762	1764	2712	1505	1627
C1	1915	1764	0928	1479	1461
C2	2170	2441	0799	2189	0814
C3	1768	1497	3098	1234	2768

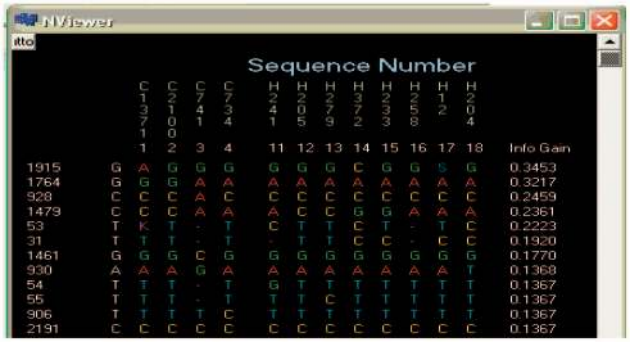

Fig. 9. The screenshot of Viewer in information gain order.

TABLE 10
The Signed Fuzzy Measure of Each Site Used in Each Data Set

Sets of sites	B	C1	C2	C3
x_1	**0.495**	**0.040**	**0.450**	**0.260**
x_2	**0.232**	0.000	0.000	0.000
x_1,x_2	0.000	0.000	**0.007**	0.000
x_3	**0.094**	**0.253**	**-0.183**	0.000
x_1,x_3	**0.175**	0.000	0.860	0.000
x_2,x_3	**-0.035**	**0.331**	0.000	0.000
x_1,x_2,x_3	0.000	0.000	0.000	**0.445**
x_4	**0.333**	0.000	**0.196**	0.000
x_1,x_4	**0.738**	0.000	**-0.604**	0.000
x_2,x_4	**0.102**	0.000	0.000	0.000
x_1,x_2,x_4	0.000	0.000	0.000	0.000
x_3,x_4	**0.252**	0.000	0.000	0.000
x_1,x_3,x_4	**0.566**	0.000	0.000	0.000
x_2,x_3,x_4	**-0.035**	0.000	0.000	0.000
x_1,x_2,x_3,x_4	0.000	0.000	0.000	0.000
x_5	**0.457**	**0.542**	**1.374**	0.000
x_1,x_5	0.000	**0.917**	**0.757**	**0.840**
x_2,x_5	0.000	**0.385**	**0.829**	**0.500**
x_1,x_2,x_5	0.000	**0.633**	**0.395**	**0.687**
x_3,x_5	0.000	**0.389**	0.000	0.000
x_1,x_3,x_5	**1.488**	**0.940**	**0.500**	**0.765**
x_2,x_3,x_5	0.000	**0.163**	**0.107**	**0.900**
x_1,x_2,x_3,x_5	0.000	**0.317**	**0.565**	**0.472**
x_4,x_5	0.000	**0.817**	**0.631**	0.000
x_1,x_4,x_5	**0.472**	**0.917**	0.000	0.000
x_2,x_4,x_5	0.000	0.000	0.000	0.000
x_1,x_2,x_4,x_5	**0.450**	0.000	**0.558**	0.000
x_3,x_4,x_5	0.000	0.000	0.000	0.000
x_1,x_3,x_4,x_5	**0.260**	**1.083**	0.000	0.000
x_2,x_3,x_4,x_5	**0.941**	**0.548**	0.000	**0.600**
X	0.000	**0.317**	**0.687**	**0.443**

For the B, C1, C2, and C3 data sets, the top five sites are used to formulate the set of linear equations. So, we obtained the solutions which have the fewest nonzeros and filtered those positions with zero. In Table 10, we show the importance and relevance of the individual sites and their interactions.

From Table 10, we can see that many sites of sequence do not take effect individually or combined with others. The nonzero sites are important for diagnosing disease. These may be helpful for bioinformatics and medical research.

5 Discussions and Future Work

In this paper, a data mining framework for DNA sequence biological data sets has been presented. It has been applied to the Hepatitis B Virus DNA data sets which are among the largest in the world and have been collected by our medical school specifically for this project. We have developed a framework for markers discovery. This framework has incorporated two algorithms, NIC and RL. Both classifiers can explicitly give the importance of the markers and their interactions and have shown good performance in cancer prediction.

Moreover, the details of the new classification method based on nonlinear integral have been presented. This method has good performance using the fuzzy measure and the nonlinear integral, due to the nonadditivity of the fuzzy measure reflecting the importance of the individual feature attributes as well as their inherent interactions. Besides the high interpretability of the Nonlinear Integrals Classifier, the experimental results have shown that it is one of the best classifiers especially in terms of sensitivity. It is very useful for preliminary diagnosis and screening test of liver cancer caused by HBV. In our model, we use GA for optimization which provides multimodal solutions containing sets of best solutions. The final confirmation experiments, like many other bioinformatics problems, need to be carried out by biochemists to identify and study the true markers. Finally, we have used a regularization method to get a solution with the fewest nonzero fuzzy measure values. It can provide some important individual and combinations of key markers of the HBV DNA sequences. We believe that this information can be helpful to do further research for biochemists.

We hypothesize that the genomic makeup of HBV affects the carcinogenic potential of the virus. In this case-control study, we have demonstrated that some genotype-specific mutations are more commonly found among HCC patients than their age and gender-matched controls. These markers can therefore be used as biomarkers to stratify the cancer risk of chronic hepatitis B patients. Our findings have been validated by independent data sets in the validation process. To confirm the biological role of these mutations, further experimental work using in situ mutagenesis of replicative HBV clones on their carcinogenicity in animal and cell line models will be required.

However, even though we have generated one of the largest data sets, the example sizes of the data sets are still small (less than 100) for each case. It is a challenge for the classifier based on nonlinear integral to avoid overtraining.

Acknowledgments

This research is partially supported by grants from the Research Grants Council of the Hong Kong SAR, China (Projects CUHK414107 and CUHK414708).

References

[1] R.P. Beasley, L.Y. Hwang, C.C. Lin, and C.S. Chien, "Hepatocellular Carcinoma and Hepatitis B Virus. A Prospective Study of 22 707 Men in Taiwan," *Lancet*, vol. 2, pp. 1129-1133, 1981.

[2] J.H. Kao, P.J. Chen, M.Y. Lai, and D.S. Chen, "Hepatitis B Genotypes Correlate with Clinical Outcome in Patients with Chronic Hepatitis B," *Gastroenterology*, vol. 118, pp. 554-559, 2000.

[3] H.L.Y. Chan et al., "Genotype C Hepatitis B Virus Infection Is Associated with an Increased Risk of Hepatocellular Carcinoma," *Gut*, vol. 53, pp. 1494-1498, 2004.

[4] H. Sumi, O. Yokosuka, N. Seki, M. Arai, F. Imazeki, T. Kurihara, T. Kanda, K. Fukai, M. Kato, and H. Saisho, "Influence of Hepatitis B Virus Genotypes on the Progression of Chronic Liver Disease," *Hepatology*, vol. 37, pp. 19-26, 2003.

[5] M.F. Yuen, Y. Tanaka, M. Mizokami, J.C. Yuen, D.K. Wong, H.J. Yuan, S.M. Sum, A.O. Chan, B.C. Wong, and C.L. Lai, "Role of Hepatitis B Virus Genotypes Ba and C, Core Promoter and Precore Mutations on Hepatocellular Carcinoma: A Case Control Study," *Carcinogenesis*, vol. 25, pp. 1593-1598, 2004.

[6] H.L.Y. Chan, C.H. Tse, E.Y.T. Ng, K.S. Leung, K.H. Lee, K.W. Tsui, and J.J.Y. Sung, "Phylogenetic, Virological and Clinical Characteristics of Genotype C Hepatitis B Virus with Tcc at Codon 15 of the Precore Region," *J. Clinical Microbiology*, vol. 44, no. 3, pp. 681-687, 2006.

[7] H.L.Y. Chan, S.K.W. Tsui, E.Y.T. NG, P.C.H. Tse, K.S. Leung, K.H. Lee, T. Mok, A. Bartholomeuz, T.C.C. Au, and J.J.Y. Song, "Epidemiological and Virological Characteristics of Two Subgroups of Genotype C Hepatitis Virus," *J. Infectious Diseases,* vol. 191, pp. 2022-2032, 2005.
[8] T. Laskus, L.-F. Wang, M.R.H. Vargas, and J. Cianciara, "Comparison of Hepatitis B Virus Core Promoter Sequences in Peripheral Blood Mononuclear Cells and Serum From Patients with Hepatitis B," *J. General Virology,* vol. 78, pp. 649-653, 1997.
[9] W.K. Keum, J.Y. Kim, J.Y. Kim, S.G. Chi, H.J. Woo, S.S. Kim, J. Ha, and I. Kang, "Heterogeneous HBV Mutants Coexist in Korean Hepatitis B Patients," *Experimental and Molecular Medicine,* vol. 30, no. 2, pp. 115-122, June 1998.
[10] R.B. Potter and S. Draghici, "A Soft Approach to Predicting HIV Drug Resistance," *Proc. Pacific Symp. Biocomputing (PSB '02),* 2002.
[11] A. Ciancio, A. Smedile, and M. Rizzetto, "Identification of HBV DNA Sequences that Are Predictive of Response to Lamivudine Therapy," *Hepatology,* vol. 39, pp. 64-73, 2004.
[12] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice," *Nucleic Acids Research,* vol. 22, pp. 4673-4680, 1994.
[13] M. Kimura, "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences," *J. Molecular Evolution,* vol. 16, pp. 111-120, 1980.
[14] N. Saitou and M. Nei, "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees," *Molecular Biology and Evolution,* vol. 4, pp. 406-425, 1987.
[15] S. Kumar, K. Tamura, and M. Nei, "MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment," *Brief Bioinformatics,* vol. 5, pp. 150-163, 2004.
[16] T. Sakamoto, Y. Tanaka, J. Simonetti, C. Osiowy, M.L. Børresen, A. Koch, F. Kurbanov, M. Sugiyama, G.Y. Minuk, B.J. McMahon, T. Joh, and M. Mizokami, "Classification of Hepatitis B Virus Genotype B into 2 Major Types Based on Characterization of a Novel Subgenotype in Arctic Indigenous Populations," *J. Infectious Diseases,* vol. 196, pp. 1487-1492, 2007.
[17] E. Orito et al., "Geographic Distribution of Hepatitis B Virus (HBV) Genotype in Patients with Chronic HBV Infection in Japan," *Hepatology,* vol. 34, pp. 590-594, 2001.
[18] F. Sugauchi, H. Kumada, H. Sakugawa, M. Komatsu, H. Niitsuma, H. Watanabe, Y. Akahane, H. Tokita, T. Kato, Y. Tanaka, E. Orito, R. Ueda, Y. Miyakawa, and M. Mizokami, "Two Subtypes of Genotype B (Ba and Bj) of Hepatitis B Virus in Japan," *Clinical Infectious Diseases,* vol. 38, pp. 1222-1228, 2004.
[19] S.M. Owyer and J.G.M. Sim, "Relationships within and between Genotypes of Hepatitis B Virus at Points Across the Genome: Footprints of Recombination in Certain Isolates," *J. General Virology,* vol. 81, pp. 379-392, 2000.
[20] H. Almuallim and T. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," *Artificial Intelligence,* vol. 69, nos. 1/2, pp. 179-305, 1994.
[21] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis,* vol. 1, no. 3, pp. 131-156, 1997.
[22] G. John, R. Kohavi, and K. Pdlwfwe, "Irrelevant Features and the Subset Selection Problem," *Proc. 11th Int'l Conf. Machine Learning,* pp. 121-129, 1994.
[23] P. Langley, "Selection of Relevant Features in Machine Learning," *Proc. AAAI Fall Symp. Relevance,* pp. 1-5, 1994.
[24] P. Pudil, J. Novovicoca, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters,* vol. 15, pp. 1119-1125, Nov. 1994.
[25] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Example Performance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 153-158, Feb. 1997.
[26] T.M. Mitchell, *Machine Learning.* The Mc-Graw-Hill Companies, Inc., 1997.
[27] C. Eugene, "Bayesian Network without Tears," *AI Magazine,* vol. 12, no. 4, pp. 50-63, 1991.
[28] D.M. Chickering, D. Heckerman, and D. Geiger, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning,* vol. 20, pp. 197-243, 1995.
[29] W. Liu, J. Cheng, and A.B. David, "An Algorithm for Bayesian Belief Network Construction from Data," *Proc. Sixth Int'l Workshop Artificial Intelligence and Statistics,* 1997.
[30] W. Banzaf, P. Nordin, R. Keller, and F. Francone, *Genetic Programming—An Introduction.* Morgan Kaufmann, 1997.
[31] A.A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery," *Advances in Evolutionary Computation,* A. Ghosh and S. Tsutsui, eds., Springer-Verlag, 2002.
[32] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain, "Dimensionality Reduction Using Genetic Algorithms," *IEEE Trans. Evolutionary Computing,* vol. 4, no. 2, pp. 164-171, July 2000.
[33] M.L. Wong and K.S. Leung, "Learning Recursive Functions from Noisy Examples Using Generic Genetic Programming," *Proc. First Ann. Conf.,* pp. 238-246, 1996.
[34] M.L. Wong and K.S. Leung, *Data Mining Using Grammar Based Genetic Programming and Applications.* Kluwer Academic Publishers, Jan. 2000.
[35] K.B. Xu, Z.Y. Wang, P.A. Heng, and K.S. Leung, "Classification by Nonlinear Integral Projections," *IEEE Trans. Fuzzy Systems,* vol. 11, no. 2, pp. 187-201, Apr. 2003.
[36] Z.Y. Wang, K.S. Leung, and J. Wang, "A Genetic Algorithm for Determining Nonadditive Set Functions in Information Fusion," *Fuzzy Sets and Systems,* vol. 102, pp. 463-469, 1999.
[37] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, 1992.
[38] S. Mika, A.J. Smola, and B. Schölkopf, "An Improved Training Algorithm for Fisher Kernel Discriminants," *Proc. Artificial Intelligence and Statistics (AISTATS '01),* T. Jaakkaola and T. Richardson, eds., pp. 98-104, 2001.
[39] M.L. Wong and K.S. Leung, "Genetic Logic Programming and Applications," *IEEE Expert,* vol. 10, no. 5, pp. 68-76, Oct. 1995.
[40] Data Mining Tools See5 and C5.0, Software, http://www.rulequest.com/see5-info.html, May 2006.
[41] SAS® Enterprise Miner (EM), http:/www.sas.com/technologies/analytics/datamining/miner/, 2009.
[42] C.C Chang and C.J. Lin, "LIBSVM: A Library for Support Vector Machines," Software, http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
[43] C. Borgelt, Bayes Classifier Induction, Software, http://fuzzy.cs.uni-magdeburg.de/~borgelt/bayes.html, 2009.
[44] H. Zhang, "The Optimality of Naive Bayes," *Proc. 17th Int'l Florida Alliance of Information and Referral Services (FLAIRS) Conf.,* 2004.
[45] C.M. Van Der Walt and E. Barnard, "Data Characteristics That Determine Classifier Performance," *Proc. 16th Ann. Symp. Pattern Recognition Assoc. of South Africa,* pp. 160-165, http://www.patternrecognition.co.za, 2006.
[46] K.S. Leung, Y.T. Ng, K.H. Lee, L.Y. Chan, K.W. Tsui, T. Mok, C.H. Tse, and J. Sung, "Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals," *Proc. Taiwan-Japan Symp. Fuzzy Systems & Innovational Computing, Third Meeting* (Keynote Speech), pp. 1-10, Aug. 2006.
[47] M.Y. Park, and T. Hastie, "L1-Regularization Path Algorithm for Generalized Linear Models," *J. Royal Statistical Soc.: Series B (Statistical Methodology),* vol. 69, no. 4, pp. 659-677, 2007.

Kwong-Sak Leung (M'77-SM'89) received the BSc (Eng.) and PhD degrees from the University of London, Queen Mary College, in 1977 and 1980, respectively. He was a senior engineer on contract R&D at ERA Technology and later joined the Central Electricity Generating Board to work on nuclear power station simulators in England. In 1985, he joined the Computer Science and Engineering Department at the Chinese University of Hong Kong, where he is currently a professor of computer science and engineering. His research interests include soft computing and bioinformatics including evolutionary computation, parallel computation, probabilistic search, information fusion and data mining, and fuzzy data and knowledge engineering. He has authored or coauthored more than 200 papers and two books in fuzzy logic and evolutionary computation. He has been a chair and member of many program and organizing committees of international conferences. He is on the Editorial Board of *Fuzzy Sets and Systems* and an associate editor of the *International Journal of Intelligent Automation and Soft Computing*. He is a senior member of the IEEE, a chartered engineer, a member of the IET and the ACM, a fellow of the HKIE, and a distinguished fellow of the HKCS in Hong Kong.

Kin Hong Lee received the degrees in computer science from the University of Manchester. Before he joined The Chinese University in 1984, he had been with Burroughs Machines in Scotland and ICL in Manchester. He is now an Associate Professor of the Computer Science & Engineering Department at the Chinese University. His current research interests include computer hardware and bioinformatics. He has published more than 60 papers in these two fields.

Jin-Feng Wang received the BS degree in computer science from Hebei Science and Technology University, Hebei, PR China, in 1999, and the MS degree in computer science from Hebei University, Hebei, PR China, in 2003. She is currently working toward the PhD degree at the Chinese University of Hong Kong. Her current research interests include nonlinear integrals and bioinformatics.

Eddie Y.T. Ng's photo and bio are not available.

Henry L.Y. Chan received the graduate degree from The Chinese University of Hong Kong and completed training at the Prince of Wales Hospital, Hong Kong. He is currently a professor in the Department of Medicine and Therapeutics, the director of Cheng Suen Man Shook Center for Hepatitis Research, and the director of Center for Liver Health of the university. He is the president of the Hong Kong Association for the Study of Liver Diseases. He is serving as an editor in the *Journal of Gastroenterology and Hepatology* and a member of the editorial board of *Clinical Gastroenterology* and *Hepatology and Alimentary Pharmacology and Therapeutics*. He has board research interest including virology of hepatitis B virus, natural history and treatment of chronic hepatitis B, liver fibrosis, hepatocellular carcinoma, and nonalcoholic fatty liver disease. He has published more than 150 peer-reviewed papers, six book chapters, and 135 conference abstracts.

Stephen K.W. Tsui is currently a professor in the Biochemistry Department of the Chinese University of Hong Kong. He is also the director of the Centre for Microbial Genomics and Proteomics and the Hong Kong Bioinformatics Centre. His major research interests are the genomics and bioinformatics of pathogenic viruses, including hepatitis B virus, SARS-coronavirus, and human immunodeficiency virus.

Tony S.K. Mok was trained at the University of Alberta and subsequently completed his fellowship in medical oncology at the Princess Margaret Hospital in Toronto. After working in a community-based oncology practice in Toronto for 7 years, he took up an academic position at the Department of Clinical Oncology at the Chinese University of Hong Kong. He has been very active in both clinical and laboratory research in the areas of lung cancer, liver cancer, and traditional Chinese medicine. He has contributed to over 100 international publications including abstracts, original articles, and book chapters. He speaks frequently at international conferences and is particularly active in China and Asia Pacific region. He is the founder of the Lung Cancer Research Group, one of the first multicenter lung cancer study group in Asia Pacific region. He is the chairman of the Hong Kong Cancer Therapy Society and executive committee member of the Chinese Society of Clinical Oncology. He has also founded the Cancer Patient Resource Center and the Cancer Information Hotline in Hong Kong. In 2002, he received the Lilly Oncology Pharmacogenomics Award in United States.

Pete Chi-Hang Tse received the MPhil degree in zoology with specialization in molecular biology from The University of Hong Kong at Hong Kong in 1997. He is now a research associate at the Institute of Digestive Diseases and Division of Gastroentology and Hepatology, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong. His main research interests are molecular virology of hepatitis B virus and viral phylogenetics.

Joseph Jao-Yiu Sung received the MBBS degree from the University of Hong Kong in 1983. He was conferred the Doctor of Philosophy by the University of Calgary and Doctor of Medicine by the Chinese University of Hong Kong in 1991 and 1997, respectively. He has held fellowships from the Royal College of Physicians of Edinburgh, London, Thailand, & Glasgow, the American College of Gastroenterology, the Royal Australian College of Physicians, the American Gastroenterological Association, and the Hong Kong College of Physicians and Academy of Medicine.