http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (10): 1749-1754, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1749.1754

Study of HBV Evolution Based on Single Nucleotide Polymorphism Using Combined Pattern Method

Ma Lei, Zhang Qi, Jia Qi-Nan, Zhang Jun, Yi Qing-Qing and He Jian-Feng
Institute of Biomedical Engineering, School of Information Engineering and Automation,
Kunming University of Science and Technology, 650500, Kunming, China

Abstract: Liver diseases such as Hepatitis B has seriously affected humans' health. According to the previous reports, gene mutation occurs as an essential factor of sustained infection. Technically, the detection of single nucleotide polymorphism loci can help study risk of infection and degree of disease evaluation. However, regular single nucleotide polymorphism selection methods such as Single-strand conformation polymorphism and Denaturing gradient gel electrophoresis are restricted by time consuming, technical difficulty and high costing. In this study, a computing method based on optimal risk and preventive pattern algorithm is applied to analyze the genetic mutation from perspective of SNPs in Hepatitis B virus Deoxyribonucleic acid sequence which illustrates the characteristics of low costing and simplified operation process. Moreover the proposed method may uncover specific unknown single nucleotide polymorphism loci from Hepatitis B virus gene sequence data. The experiment releases 18 potential single nucleotide polymorphisms locus that demonstrate the transition from Acute to Chronic, where five of them have been reported in previous research, the remaining 13 are potential locus. In addition, 3 potential single nucleotide polymorphisms locus that demonstrated the transition to Cirrhosis are discovered. It may provide valuable reference for the clinic prevention and diagnostics.

Key words: Data mining, genetic loci, single nucleotide polymorphism, HBV

INTRODUCTION

Hepatitis B Virus (HBV) is an infectious disease that affects millions of people globally. It can lead to chronic hepatitis and particularly prevalent in developing countries. Currently, more than 240 million people are estimated to be chronically infected with hepatitis B worldwide (WHO, 2012). In 2010, Hepatitis B and hepatitis C have totally caused 1.4 million deaths, including deaths from acute infection, cirrhosis and liver cancer (Lozano *et al*., 2013; WHO, 2013).

Until now, countless efforts into Hepatitis B cure and precaution have been devoted. Consequently some explorations release that the detection of specific Single Nucleotide Polymorphisms (SNPs) loci can certainly help to predict the risk of individual HBV infection and determine the degree of disease evolution after being infected, from which pertinent treatment and prevention can be beneficial (Zheng *et al*., 2010; Lau, 2007; Banoub and Limbach, 2010; Ng, 2008).

SNP means the polymorphism of DNA sequence caused by mutation of single nucleotide in a certain region of chromosome genome; over 1% individuals in an appointed biotic population can be involved in the mutation. The format basically includes transformation, transpose, insertion and deletion of single nucleotide. In the analysis of hereditism, SNP is treated as a class of genetic tab broadly applied and the characteristic of high density, representativeness, genetic stabilization and automatic analysis actuate it to be the third generation of hereditary label following the restriction fragment length polymorphism(RFLP) and microsatellite technologies (Barreiro *et al*., 2008; Varela and Amos, 2010; Fareed and Afzal, 2012).

Nowadays, the methods of SNPs loci detection normally include Single-strand conformation polymorphism (SSCP) (Kubo *et al*., 2009), Denaturing gradient gel electrophoresis (DGGE) (Drabovich and Krylov, 2006) and some others (RFLP Website), (Den Dunnen and Antonarakis, 2000; Sapolsky *et al*., 1999; Gross *et al*., 1999). However, the above experimental methods are limited by long time consumption, complex operations, technical difficulty and high costs. Therefore, in this study, a method based on optimal risk and preventive pattern is applied for SNPs loci selection, in order to provide a convenient and simplified tool for

Corresponding Author: He Jian-Feng, Institute of Biomedical Engineering, School of Information Engineering and Automation, Kunming University of Science and Technology, 650500, Kunming, China

study of HBV gene mutation during the transformation from acute to chronic carrier, then to cirrhosis.

MATERIALS AND METHODS

Extraction of feature based on information entropy: During the collection of gene data samples, some redundant information could be produced due to multi-factors from outside. Therefore, an extraction of feature information must be applied to filter redundant data that regularly represents non-disease gene or non-risk gene, while, the remaining feature information can be better applied for diagnosis, prevention and treatment.

In this study, a method of feature extraction based on entropy is going to be introduced. With information gain in DNA sequences, the criterion of feature information selection is applied to judge the information brought to a target attribute by feature gene on the loci, the more info increases, the more indispensable a feature gene is. To a certain feature gene, amount of information will vary dramatically when it exists or does not exist on the loci, D-value of variation is the amount of information brought to the loci by feature gene, it is also named entropy (Cover and Thomas, 2012) calculated by:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \quad (1)$$

The more possible variation of X is, the more information can be carried by X and the larger value of entropy is. Therefore, the info gain brought by cluster C or classification C by feature info T is represented by Eq. 2:

$$IG(T) = H(C) - H(C|T) \quad (2)$$

The H (C|T) contains two conditions respectively: Firstly, the feature T labeled as t occurs. The other is that feature T labeled as "t"' doesn't appear. The equation for H (C|T) calculation is shown in Eq. 3:

$$H(C|T) = P(t)\,H(C|t) + P(t')\,H(C|t') \quad (3)$$

The info gain equation of feature and classification can be derived from the Eq. 2. This method is implemented and achieved on WEKA platform.

The high complexity and large scale of biological data could obviously lead to the difficulty for data mining. Therefore, the gene data must be transformed to the format required during the research. The way of mapping is demonstrated in Fig. 1.

Optimal risk and preventive pattern: As an algorithm of association rules, the core idea of Mining Optimal Risk pattErnsets (MORE) (Li *et al*., 2009) is to apply local support (lsupp) (Ohsaki *et al*., 2004) for frequent pattern discovery, then the Relative Risk (RR) (Triola and Triola, 2005) that is regularly occurred in epidemic as a metric index is utilized to generate optimal risk and preventive pattern. Pattern represents the combination of nucleotides from different locus. For example, (At. 1 = A, At. 51 = C, At. 45 = T) can be regarded as a pattern. The basic concept and calculation of lsupp and RR are briefly introduced below.

For some real clinic datasets, the data amount is pretty huge and the classifications tabbed on positive or negative keep serious imbalance. The rate of infection is much less than that of non-infection, therefore, the lsupp is applied as a support index of risk pattern which can filter non-infection samples. It can be represented as a value of ratio that equals to the probability of simultaneous occurrence of pattern P and a, divided by the probability of alone occurrence of pattern a. It is represented by Eq. 4:

$$lsupp(p \rightarrow a) = \frac{supp(p \rightarrow a)}{supp(a)} \quad (4)$$

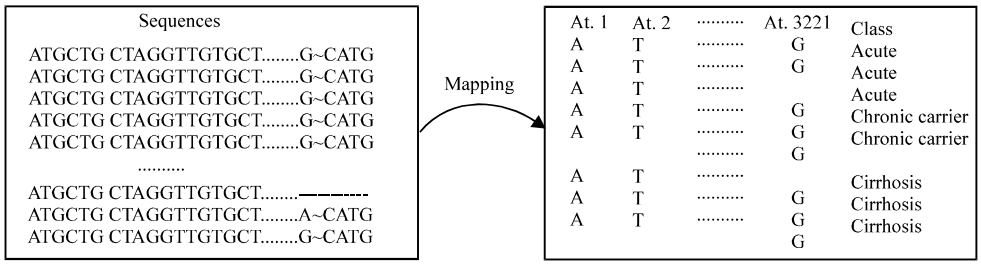

Fig. 1: Mapping of the gene sequence data to the required format. Each sequence corresponds to the No. of sample, attribute 1 to attribute 3221 represent the basic group related to certain loci in each sequence, class represents the corresponded class attribute of each sequence, three types of liver disease are denoted in the right panel

Table 1: Results and possibility of patterns

Parameters	Abnormal(a)	Normal(n)	Total
P	Supp (p,a)	Supp (p,n)	Supp (p)
¬P	Supp (¬p,a)	Supp (¬p,n)	Supp (¬p)
Total	Supp (a)	Supp (n)	1

In Eq. 4, supp (p→a) means a support index of pattern P, it also represents the probability of simultaneous occurrence of pattern P and a. lsupp satisfies the requirement of anti-monotonicity rule which illustrates that the support index of a superior set is equal or less than the index of any subset. The satisfaction of anti-monotonicity rule is the basic law for an optimal risk and preventive pattern mining. In this study, if the lsupp of a pattern is larger than the allocated threshold, the pattern is recognized as frequent.

As mentioned above, RR equals to an incidence rate of a certain group divided by the rate of uncertain people. The threshold of RR needs to be selected. For instance, if the threshold value is 1 (RR = 1), it denotes that a pattern is neither risk nor preventive pattern, there is no influence on the infection of patient; If RR>1, the pattern is appointed as risk, the attribute of this pattern is risk factor; If RR<1, the pattern is preventive, the attribute is preventive factor. RR can be calculated by Eq. 5. All the risk patterns compose risk set and the preventive set is comprised of all preventive patterns:

$$\begin{aligned} RR(p \to a) &= \frac{supp(a|p)}{supp(a|\neg p)} \\ &= \frac{supp(pa)/supp(p)}{supp(\neg pa)/supp(\neg p)} \\ &= \frac{supp(pa)supp(\neg p)}{supp(\neg pa)supp(p)} \end{aligned} \quad (5)$$

In Table 1, ¬p represents all records of non-occurrence with pattern P. Supp (¬p) means all records without pattern P, ¬p shows all records including a but without pattern P.

Variation analysis of feature attribute based on optimal risk and preventive pattern sets: Occasionally, contradictions are caused by the disease related or non-related patterns mined by optimal risk and preventive pattern. In this way, an attribute may occur in both the optimal risk and preventive sets. It may also lead to clinical diagnostic errors. With an intention to solve the problem, we assume each feature attribute to be within optimal risk and has it's independent preventive pattern is independent. Next, a risk set is generated by feature attribute set of optimal risk pattern. Meanwhile, a preventive set is generated by feature attribute set of optimal preventive pattern. Then, the frequency of feature attribute of risk and preventive sets is counted. The public feature properties with equal frequent value have been deleted. Moreover, frequent sets of optimal risk and preventive are generated. The feature attribute concentrated on optimal risk frequent set is treated as risk factor, it represents that the patient who carries the gene is under the risk of infection. A feature attribute focused on optimal preventive frequent set is labeled as the preventive factor; it means that the patient who carries the gene is without the risk of infection. Finally, normalization is applied on the frequency of each feature attribute, the weight value of feature attribute is produced, and sum of weight is 100. We present a more clear explanation through an example by the next.

Here is the example. Suppose we have six risk patterns (the relative risk threshold: 2.0) as follows:

- {R1 = A, R2 = G, R4 = A} (RR = 10)
- {R1 = A, R2 = G, R3 = T}(RR = 4.3)
- {R2 = G, R3 = T, R5 = C, R6 = C} (RR = 3.7)
- {R2 = C, R4 = A, R5 = G} (RR = 1.0)
- {R1 = A, R2 = C, R5 = G, R6 = C} (RR = 1.3)
- {R1 = A, R4 = A, R6 = C} (RR = 0.1)

These patterns involve six patterns and eight attributes: R1 = A, R2 = G, R2 = C, R3 = T, R4 = A, R5 = G, R5 = C, R6 = C. As we discussed earlier, only the first three are risk patterns the left other three are preventive patterns. Risk patterns involve six attributes: R1 = A, R2 = G, R3 = T, R4 = A, R5 = C, R6 = C. Preventive patterns involve six attributes: R1 = A, R2 = C, R4 = A, R5 =G, R6 = C. Table 2 shows the weight of each attribute in risk patterns and preventive patterns.

As we can see, the weight of R1 = A in risk patterns and preventive patterns are equal. So, we delete it. The weights of R4 = A, R6 = C in preventive patterns is bigger than these in risk patterns. So, we delete R4 = A, R6 = C in risk patterns. At last, these remain attributes are selected to generate optimal risk and preventive sets. In this example the optimal risk sets are R2 = G, R3 = T, R5 = C. The optimal risk sets are R2 = C, R4 = A, R5 = G, R6 = C.

RESULTS AND DISCUSSION

Data source for the experiment: Gene sequence data applied in this study is downloaded from Genbank (Genbank Website). Each sequence contains 3221 locus which can be applied for the analysis of molecular evolution and mutation pattern of Hepatitis B. Forty selected s ample sets of sequence include FJ349205-FJ349241, EU859930-EU859937 and EU859939-EU859956.

Table 2: Optimal risk patterns

Pattern	Length	RR	Attribute
1	1	3.6667	1773
2	1	3.6667	49

Lsupp: 0.05, Pattern length: 6, Threshold of feature attribute: 0.20, RR threshold: 1, selected optimal risk patterns are generated from HBV sequence

Table 3: Optimal preventive patterns

Pattern	Length	RR	Attribute
1	4	0.0000	1028
2	4	0.0000	902
			507
			301
			1231
			1028
			507
			301

Lsupp: 0.05, Pattern length: 6, Threshold of feature attribute: 0.20, RR threshold: 1, selected optimal risk patterns are generated from HBV sequence

Among them 9 samples are acute, 22 samples are chronic carriers and 9 samples are cirrhosis. The objective of this experiment is to uncover SNP locus of HBV from the sequence and data processing is carried on according to the law that each vertical column in HBV sequence maps a feature attribute. The data types of HBV sequence include acute, chronic carrier and cirrhosis, respectively.

Generation of optimal risk and preventive pattern: The method mentioned in materials and methods is applied to study the relation between acute, chronic carrier and cirrhosis. The length of experimental HBV sequence is 3221bp after multi-alignment which directly map on 3221 vertical column. Numerous of experiments are proceeded to select the optimal pattern length and threshold of RR. Lastly, a preferred solution is chosen; pattern length is 6, threshold of feature attribute is set to 0.20, RR threshold equals to 1. Under this condition, the experiment retrieves 400 optimal risk and preventive patterns which contain 36 optimal risk patterns and 364 preventive patterns. Due to the limitation of this study, only the representative optimal risk patterns in Table 2 and preventive patterns in Table 3 are listed.

To further explain the experimental results in Table 2 and 3, we take Pattern 1 in optimal risk pattern as an example. Length = 1 means that the length of pattern is 1, it represents that this pattern contains one feature attribute, RR = 3.6667 illustrates that the value of RR is 3.6667. Risk set covers a pattern with relatively high possibility of mutation, Pattern 1 in risk set is used as an example for explanation. When basic group on the 1773th loci is A, the possibility of mutation is relatively high due to RR = 3.6667. If it's not A, the possibility of mutation is relatively low. Conversely, an optimal preventive set denotes a pattern with possibility that the mutation may not occur. For instance, Pattern 1 in preventive set contains four attributes that correspond to the basic group of four locus, RR = 0.0000 means that there will be almost no mutation when this pattern occurs.

Variation analysis of HBV sequence: In the experiment, we suppose that each feature attribute of patterns is independent, the optimal risk and preventive weight of feature attribute of HBV sequence can be shown in Table 4 and 5, respectively. Each weight initiates from its corresponding ratio in optimal risk and preventive sets, it can be applied to judge the level of importance of each related feature attribute, meanwhile the hazard and preventability of disease related to a creation feature attribute can be recognized.

In Table 4, feature attribute attribute 2531 = T occurs in the optimal risk set, attribute 2531 corresponds to the 2531th loci of basic group in a HBV sequence. The risk weight is 10.4427 and represents the maximum risk weight in optimal risk set which releases that the possibility of transfer to chronic carrier from acute is topmost among all feature attributes when the 2531th loci is T which is regarded as a determinant of mutation. Attribute 2567 appears in optimal preventive set, the preventive weight is 8.9241 which reveals that the possibility of Non-transfer from acute to chronic carrier is relatively high, when the 2567th loci is T which is treated as a determinant of mutation prevention. If a certain attribute displays in optimal risk and preventive set simultaneously, it is a potential SNPs loci that determines transfer from acute to chronic carrier.

Accordingly, based on optimal risk weight set, this experiment releases 18 potential SNPs locus related to the transformation from acute to chronic carrier, four of them (nt1655, nt1753, nt1762, nt1764) belong to mutation of basic group replacement, one (nt1896) is regarded as deletion. The mentioned 5 SNPs locus have been reported by previous research (Yuan *et al*., 2009; Wang *et al*., 2000). The remaining 13 SNPs appear as new found locus, among them, nt2576 is replacement mutation and others are deletion.

As it can be seen from the optimal risk weight sets in Table 5, three potential SNP locus related to the transformation from chronic carrier to cirrhosis are filtered out, two of them (nt2159' nt902) belong to mutation of basic group replacement, the third one (nt507) is regarded as deletion.

Table 4: Descending ranking of frequency of each feature attribute in the optimal risk and preventive sets generated from the data of acute and chronic HBV carrier

Optimal risk and preventive pattern (supp = 0.43, info gain>0.15, pattern length = 7)					
Attribute in risk pattern			Attribute in preventive pattern		
Risk factor	Loci	Weight	Preventive factor	Loci	Weight
Attribute 2531 = T	nt2531	10.4427	Attribute 2567 = T	nt2567	8.9241
Attribute 1896 = G	nt1896	9.2698	Attribute 1773 = A	nt2567	8.0144
Attribute 1588 = A	nt1588	7.3401	Attribute 3090 = A	nt3090	7.0107
Attribute 2689 = T	nt2689	6.8104	Attribute 1421 = G	nt1421	5.756
Attribute 1320 = C	nt1320	6.5834	Attribute 1657 = C	nt1657	5.458
Attribute 1655 = T	nt1655	6.3942	Attribute 49 = G	nt49	5.3795
Attribute 1762 = T	nt1762	6.3564	Attribute 1764 = G	nt1764	5.0502
Attribute 681 = C	nt681	5.1457	Attribute 2576 = T	nt2576	4.5483
Attribute 3073 = T	nt3073	4.7673	Attribute 1999 = A	nt1999	4.3444
Attribute 3094 = C	nt3094	4.5781	Attribute 1334 = C	nt1334	4.313
Attribute 1764 = A	nt1764	4.0484	Attribute 917 = G	nt917	4.313
Attribute 1753 = T	nt1753	4.0484	Attribute 3129 = A	nt3129	4.2033
Attribute 3094 = C	nt3094	3.3674	Attribute 1762 = T	nt1762	4.1248
Attribute 2819 = A	nt2819	3.1025	Attribute 600 = C	nt600	4.1248
Attribute 2687 = G	nt2687	2.9512	Attribute 2260 = A	nt2260	3.9837
Attribute 2576 = G	nt2576	2.6485	Attribute 8 = A	nt8	3.3407
Attribute 2588 = C	nt2588	2.4593	Attribute 2717 = T	nt2717	3.2465
Attribute 681 = C	nt681	2.1188	Attribute 9 = A	nt9	3.2152
			Attribute 1655 = C	nt1655	3.2152
			Attribute 513 = C	nt513	3.074
			Attribute 1753 = A	nt1753	2.3683
			Attribute 896 = T	nt896	1.9918

Table 5: Descending ranking of frequency of each feature attribute in the optimal risk and preventive sets generated from the data of chronic HBV carrier and cirrhosis

Optimal risk and preventive pattern (supp = 0.05, info gain>0.2, pattern length = 6)					
Attribute in risk pattern			Attribute in preventive pattern		
Risk factor	Loci	Weight	Preventive factor	Loci	Weight
Attribute 2159 = G	nt2159	50	Attribute301 = A	nt301	12.8205
Attribute 902 = T	nt902	25	Attribute881 = T	nt881	12.0761
Attribute 507 = T	nt507	25	Attribute1028 = T	nt1028	11.4971
			Attribute902 = C	nt902	10.0083
			Attribute956 = A	nt956	9.5947
			Attribute576 = C	nt576	8.6849
			Attribute1370 = C	nt1370	8.354
			Attribute1231 = T	nt1231	7.8577
			Attribute3081 = A	nt3081	7.775
			Attribute2973 = C	nt2973	7.775
			Attribute2159 =T	nt2159	1.6543

CONCLUSION

In recent years, methods of SNP locus detection are broadly studied and some can provide efficacious information for the prevention of transformation from HBV acute to chronic carrier and cirrhosis. Worldwide experts have continuously proposed many techniques on this. However, most of them are highly relied on the support of high price equipment and academic staff. This study proposes a method by applying optimal risk and preventive pattern algorithm to study the detection of SNP locus which can effectively and inexpensively accomplish the filtration of SNPs locus from large scale of gene sequence database. From 40 selected HBV sequence, the experiment releases 18 potential SNPs locus related to the transformation from acute to chronic carrier, 5 of them have been reported by previous research, the remaining 13 SNPs importantly appears as new found locus and also our results discovered 3 SNPs locus that potentially represents the transformation from chronic carrier to cirrhosis. Compared with the traditional methods, the proposed method takes advantage in universal property and relevance ratio, only common PC and software are required, In addition, the proposed method can possibly be a regular method to be applied in clinic and research for the diagnosis, prevention and treatment of HBV or other diseases.

ACKNOWLEDGMENTS

This work was financially supported by China National Natural Science funding (No. 11265007) and Science Research Foundation for the Overseas Chinese Scholars, State Education Ministry (No. 2010-1561).

REFERENCES

Banoub, J.H. and P.A. Limbach, 2010. Mass Spectrometry of Nucleosides and Nucleic Acids. Taylor and Francis, Oxford, UK.

Barreiro, L.B., G. Laval, H. Quach, E. Patin and L. Quintana-Murci, 2008. Natural selection has driven population differentiation in modern humans. Nat. Genet., 40: 340-345.

Cover, T.M. and J.A. Thomas, 2012. Elements of Information Theory. Wiley-Interscience, New York.

Den Dunnen, J.T. and S.E. Antonarakis, 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Human Mutat., 15: 7-12.

Drabovich, A.P. and S.N. Krylov, 2006. Identification of base pairs in single-nucleotide polymorphisms by MutS protein-mediated capillary electrophoresis. Anal. Chem., 78: 2035-2038.

Fareed, M. and M. Afzal, 2012. Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service. Egyp. J. Med. Human Genet., 14: 123-134.

Gross, E., N. Arnold, J. Goette, U. Schwarz-Boeger and M. Kiechle, 1999. A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. Human Genet., 105: 72-78.

Kubo, K.S., R.M. Stuart, J. Freitas-Astua, R. Antonioli-Luizon and E.C. Locali-Fabris *et al*., 2009. Evaluation of the genetic variability of orchid fleck virus by single-strand conformational polymorphism analysis and nucleotide sequencing of a fragment from the nucleocapsid gene. Arch. Virol., 154: 1009-1014.

Lau, C.C., 2007. Hepatitis B Virus and Single Nucleotide Polymorphisms. Baptist University, Hong Kong.

Li, J., A.W.C. Fu and P. Fahey, 2009. Efficient discovery of risk patterns in medical data. Artif. Intell. Med., 45: 77-89.

Lozano, R., M. Naghavi, K. Foreman, S. Lim and K. Shibuya *et al*., 2013. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the global burden of disease study 2010. Lancet, 380: 2095-2128.

Ng, L.J., 2008. Hepatitis B and C Infection, Single Nucleotide Polymorphisms of the Inflammatory Pathway and the Risk of Non-Hodgkin's Lymphoma. ProQuest, USA., ISBN-13: 9781109019902, Pages: 161.

Ohsaki, M., S. Kitaguchi, K. Okamoto, H. Yokoi and T. Yamaguchi, 2004. Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis. In: Knowledge Discovery in Databases: PKDD 2004, Boulicaut, J.F., F. Esposito, F. Giannotti and D. Pedreschi (Eds.). Springer, Berlin, Germany, pp: 362-373.

Sapolsky, R.J., L. Hsie, A. Berno, G. Ghandour, M. Mittmann and J.B. Fan, 1999. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. Genet. Anal. Biomol. Eng., 14: 187-192.

Triola, M.M. and M.F. Triola, 2005. Biostatistics for the Biological and Health Sciences. 2nd Edn., Addison-Wesley, Boston, USA.

Varela, M.A. and W. Amos, 2010. Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence. Genomics, 95: 151-159.

WHO, 2012. Prevention and control of viral hepatitis infection: Framework for global action. World Health Orgnization, Geneva, Switzerland.

WHO, 2013. Global policy report on the prevention and control of viral hepatitis in WHO member states. World Health Orgnization, Geneva, Switzerland.

Wang, Y., H. Liu, Q. Zhou and X. Li, 2000. Analysis of point mutation in site 1896 of HBV precore and its detection in the tissues and serum of HCC patients. World J. Gastroenterol., 6: 395-397.

Yuan, J.M., A. Ambinder, Y. Fan, Y.T. Gao, C.Y. Mimi and J.D. Groopman, 2009. Prospective evaluation of hepatitis B 1762T/1764A mutations on hepatocellular carcinoma development in Shanghai, China. Cancer Epidemiol. Biomarkers Prev., 18: 590-594.

Zheng, L., D. Li, F. Wang, H. Wu and X. Li *et al*., 2010. Association between hepatitis B viral burden in chronic infection and a functional single nucleotide polymorphism of the PDCD1 gene. J. Clin. Immunol., 30: 855-860.