B 型肝炎基因序列特異點分析之研究

孫光天 [a]*，劉昱均 [a]，劉威良 [a]，劉紋君 [b]，張定宗 [c]

國立臺南大學 資訊教育研究所 [a]

成功大學基礎醫學研究所 [b]　成功大學分子醫學研究所 [c]

* ktsun@mail.nutn.edu.tw

摘要

本研究係由成功大學醫院臨床資料 11 位病患為對象，應用資料探勘（Data Mining）技術從 B 型肝炎病患的四個有意義的抽血時間中，分析基因序列位置的特異點與「B 型肝炎 e 抗原是否轉態」、「B 型肝炎基因型」和「肝發炎指數是否正常」之間的關係。本研究先應用序列式樣本探勘技術出 17 條規則，再透過決策樹從這些規則中找出最佳分類規則。研究結果顯示 B 型肝炎基因序列特異點位置 216 以及 1574 可分類轉態與否；以特異點位置 216 可分類基因型 B 或 C、以特異點位置 1930 以及 2076 可分類肝發炎指數正常與否，其辨識正確率皆達 100%。

關鍵詞：資料探勘、序列式樣本、B 型肝炎病毒

壹、緒論

B 型肝炎病毒核苷酸序列全長為 3215。涵蓋四個區，稱開放讀碼區，分別為 S (surface protein gene) 表面蛋白基因、C (core protein gene) 核心蛋白基因、P (polymerase protein gene) 聚合酶蛋白基因，以及 X 基因（X gene）。這四區可再細分為七個基因片段，各基因片段在病毒複製過程中各司其職，發揮其主要的特性[2]，說明如下：

(1) S 區：包括 PreS1 基因片段、PreS2 基因片段與 S 基因片段。PreS1 基因片段是 B 型肝炎病毒不同亞型的變異和 B 型肝炎疾病的活躍複製有關；PreS2 基因片段是編碼聚合人血清白蛋白受體的位點和 B 型肝炎病毒的活躍複製亦有關。

(2) C 區：包括 PreCore 基因片段與 C 基因片段。C 基因片段前連接 PreCore 基因片段，僅含 C 基因片段可表達 B 型肝炎病毒核心抗原。核心抗原鄰近 B 型肝炎病毒基因長鏈的缺口處，是 B 型肝炎病毒與受感染宿主肝細胞中去氧核糖核酸相整合的部份。

(3) P 區：P 基因片段與 S、C、X 基因片段相重疊。在病毒複製中具有逆轉錄酶的活性。

(4) X 區：在感染 B 型肝炎病毒的患者血清中，如果 X 抗原持續陽性，可能顯示患者的 B 型肝炎由急性向慢性轉化；而抗 X 抗體的持續陽性則說明 B 型肝炎感染已經慢性化。由於抗 X 抗體在表面抗原陽性的原發性肝細胞癌患者中檢出率甚高，所以認為一旦 X 基因片段隨 B 型肝炎病毒整合到宿主肝細胞的去氧核糖核酸中後，可能是誘發肝細胞癌的潛在基礎。

目前科學家發現許多 B 型肝炎病毒的突變株，突變的位置主要集中在某些區域，尤其是在 basal core promoter (BCP)、PreCore (PC)和 Pre-S 等區域。研究顯示在基因型 A 和 C 型病毒株，易出現 BCP 突變株，而 PC 突變株的產生則侷限於基因型 B、C 和 D 型。類似地，基因型 B 和 C 相比，pre-S 缺損之突變株較常出現在感染基因型 C 之 B 型肝炎病患身上[7]。國家衛生研究院研究發現，B 肝帶原者的時間愈久，肝臟體出現 pre-S 突變株比率與演變肝癌機會也愈大。歷時十年追蹤 200 名 B 肝帶原者，經過 10 年時間，發現其中 25 名有 pre-S 突變株的帶原者，14 人演變成肝癌，比率達 56%，未帶此突變株的 175 名帶原者中，罹患肝癌的比率僅為 28% [1]。

以往 B 型肝炎病毒序列位置突變的研究都是針對不同病人單一 B 型肝炎病毒序列突變位置找出相關性。但實際上序列突變不只會突變一次，在長時間的觀察下所突變的次數和突變類型都會呈現不同的變化。由於這樣的序列變化存在著時間上的關係，因此在本研究透過四個不同時間的抽血點的序列，並使用資料探勘技術中的循序樣式之探勘，找出序列的循序規則，再

利用決策樹歸納出各病患在 e 抗原血清是否轉態及不同抽血參數值的循序式序列表現。

貳、實驗方法

一、研究對象

本研究資料來源，由「成功大學醫學院分子醫學研究所張定宗教授實驗室」所提供的臨床病人資料。病患共 11 位，平均每個病患記錄 18 個月抽血之血液參數。其中男性病患有 7 位、女性病患有 4 位，平均年齡為 35 歲。其它關於 B 型肝炎病毒特徵如表 1。

表 1、臨床病人資料

B 型肝炎病毒特徵	病患數（人）
轉態 / 不轉態	3 / 8
基因型 B / 基因型 C	3 / 8
肝發炎指數正常 / 肝發炎指數異常	3 / 8

註：肝發炎指數正常為肝發炎指數比正常值上升倍數介於 2 至 5。

二、序列式樣本

(一)、序列式樣本說明

Agrawal and Srikant 於 1995 年最早提出序列樣式探勘法（sequential pattern scheme），其主要目的是從序列資料庫中，找出大多數人「頻繁」(frequent) 序列樣式（sequential pattern），其中頻繁指的是大於最低支持度的序列樣式，而最低支持度（minimum support）則定義為序列樣式探勘中，分析者事先給定的門檻值（threshold），而且序列要高於門檻值，才能稱為頻繁序列樣式[3,4,5]。

假設資料庫是一群顧客交易的集合，其中每筆交易 T 由顧客編號（CID）、交易時間（TID）和項目集合（a set of items）所組成。一個項目集 (Itemset) ，以（x1, x2,…, xm） 來表示，為一個非空項目集合。一個序列 s，以〈a1, a2, …, an〉表示是一個項目集的有序集合。每個 aj 是一個項目集，它被稱做是序列的一個元素（element)。一個項目只會出現在同一元素中一次，但能夠出現在同一序列不同元素中多次。一個 k-序列 (k-sequence) 由 k 個項目所組成 (也就是長度為 k 的序列) 。

序列 s1 為〈a1, a2,…, an〉和序列 s2 為〈b1, b2,…, bm〉，假如它們之間存在著整數 i1＜ i2＜…＜ in 使得 a1⊆bi1, a2⊆bi2 …, an⊆bin，則稱序列 s1 是序列 s2 的子序列 (subsequence) 。例如〈(c), (b, e), (b)〉是〈(c), (b), (a), (b, e), (b, f)〉的一個子序列，因為（c）⊆（c），（b, e）⊆（b, e）和（b）⊆（b, f)。然而〈(b)，(e)〉不是屬於〈(b, e)〉的子序列[5]。

如果最小支持度為 s，序列資料庫中至少 s 個序列含有 n-sequence 的序列，則將此 n-sequence 定義為 large n-sequence。候選序列（candidate sequence）則定義尚未確定是否滿足最小支持度的序列可能成為 large n-sequence。以 Lk 表示 large k-sequences 所形成的集合；以 C_k 代表 candidate k-sequences 所形成的集合。

(二)、資料處理流程

本研究將整個資料處理流程分為五個步驟，其步驟說明如下：

步驟一、篩選抽血點

首先從 11 位病患之抽血參數中的肝發炎指數（ALT）及病毒量（HBV DNA)，計算並找出本研究所需要之四個時間點的病毒序列。其四個時間點的定義如下圖 1 所示。

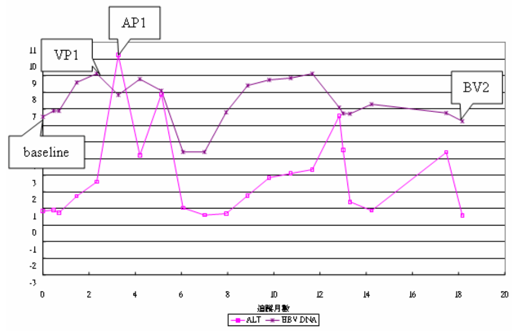

圖 1、四個抽血點分佈位置

Baseline：為病人第一個的抽血點。

VP1 (Viral Peak 1)：B 型肝炎病毒量堀起的第一個波峰。

AP1 (ALT Peak 1)：肝發炎的第一個波峰。

BV2：病人最後一個抽血點 (Endpoint)。

步驟二、序列比對

序列比對的目的在於將 11 位病患之 44 株病毒序列排列為等長序列。方法為將符合條件的病人資料先轉成 Fasta 格式，再透過 Clustal X 進行比對。。

Fasta 格式為以大於 (>) 的符號開頭作為第一欄位，所有序列長度每行不大於 80 字元。如下圖 2 所示。

```
>L200-VP1
ATGGGAGGTTGGTCTTCCAAACCTCGACAAGGCATGGGGACGAATCTTTCTGTTCCCAATCCTCTGGGATTCTTTCCCGA
TCACCAGTTGGATCCTGCGTTCGGAGCCAACTCAAACAATCCAGATTGGGACTTCAACCCCAACAAGGATCACTGGCCAG
AGGCAAATCAGGTAGGAGTGGGAGCATTCGGGCCAGGGTTCACCCCACCACACGGCAGTCTTTTGGGGTGGAGCCCTCAG
GCTCAGGGCATATTGACAACAGTGCCAGCAGCACCTCCTCCTGCCTCCACCAATCGGCAGTCAGGAAGACAGCCTACTCC
CATCTCTCCACCTCTAAGAGACAGTCATCCTCAGGCC

>L200-AP1
ATGGGAGGTTGGTCTTCCAAACCTCGACAAGGCATGGGGACGAATCTTTCTGTTCCCAATCCTCTGGGATTCTTTCCCGA
TCACCAGTTGGATCCTGCGTTCGGAGCCAACTCAAACAATCCAGATTGGGACTTCAACCCCAACAAGGATCACTGGCCAG
AGGCAAATCAGGTAGGAGTGGGAGCATTCGGGCCAGGGTTCACCCCACCACACGGCAGTCTTTTGGGGTGGAGCCCTCAG
GCTCAGGGCATATTGACAACAGTGCCAGCAGCACCTCCTCCTGCCTCCACCAATCGGCAGTCAGGAAGACAGCCTACTCC
CATCTCTCCACCTCTAAGAGACAGTCATCCTCAGGCC

>L200-BV2
ATGGGAGGTTGGTCTTCCAAACCTCGACAAGGCATGGGGACGAATCTTTCTGTTCCCAATCCTCTGGGATTCTTTCCCGA
TCACCAGTTGGATCCTGCGTTCGGAGCCAACTCAAACAATCCAGATTGGGACTTCAACCCCAACAAGGATCACTGGCCAG
AGGCAAATCAGGTAGGAGTGGGAGCATTCGGGCCAGGGTTCACCCCACCACACGGCAGTCTTTTGGGGTGGAGCCCTCAG
GCTCAGGGCATATTGACAACAGTGCCAGCAGCACCTCCTCCTGCCTCCACCAATCGGCAGTCAGGAAGACAGCCTACTCC
CATCTCTCCACCTCTAAGAGACAGTCATCCTCAGGCC
```

圖 2、Fasta 格式範例

符合格式之後再透過序列比對軟體 — Clustal X 進行比對(如圖 3 所示)。

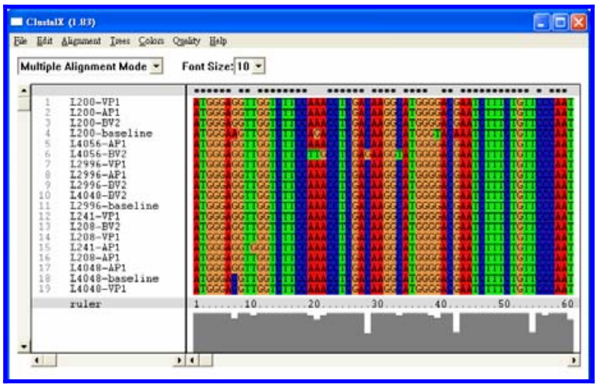

圖 3、序列比對軟體畫面

步驟三、尋找特異點

將比對後序列，透過 Gene Doc 進行編輯以利特異點的擷取。而特異點的定義為，在同一個位置上之序列發生不同的鹼基表現則本研究稱之為特異點。圖 4 中 Gene Doc 程式尋找特異點情形，縱向位置全部以「‧」符號呈現表示此位置未發生過突變，則不是特異點需過濾掉，而縱向位置有多數「‧」符號夾雜 A、T、C、G 等鹼基表示該點為突變產生，則此位置為特異點。

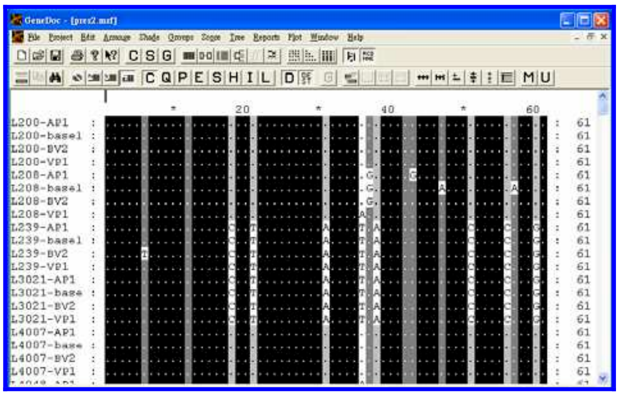

圖 4、 病人特異點擷取

步驟四、序列式樣本探勘

接下來將尋找出的特異點，利用序列式樣本探勘找出最常出現的序列樣式。表 2 為以門檻值設定為 2 所探勘出序列的規則。

表 2、基因片段規則

特異點	抽血點				病患編號										
Position	baseline	VP1	AP1	BV2	L200	L208	L221	L239	L241	L2996	L3021	L4007	L4035	L4048	L4056
1896	G	G	A	G				*				*			
1930	G	G	T	G			*			*					
2076	T	A	A	T					*	*					
2143	A	A	T	A			*			*					
2149	T	C	T	T		*			*						
2159	C	G	G	G							*		*		
2259	A	A	G	A		*						*			
2446	T	C	C	C							*		*		
25	G	A	G	G		*			*						
216	C	T	T	T	*	*			*	*		*		*	*
389	C	C	T	C		*	*								
1499	T	A	A	T					*	*					
1574	C	A	A	C					*	*					
1574	A	A	A	A	*		*	*				*	*	*	
1613	G	G	A	G	*									*	
1703	G	G	G	A								*			*
1703	G	A	A	G					*	*					

表 2 以特異點位置為 1896 為例，在抽血點 baseline 時，其鹼基表現為 G、VP1 時為 G、AP1 時為 A、BV2 時為 G。符合此特異點病患有 2 位，分別為 L239 以及 L4007。

步驟五、歸納分類

將探勘結果轉換成決策樹輸入格式並設定特異點為輸入屬性、轉態、基因型和 ALT 上升倍數為目標值。以表 3 為例，各病患符合特異點位置則設定為 1，不符合則為 0。其中病患 L200、L4007、L4048 皆是屬於轉態，目標值設定為 YES 其餘病患設定為 NO。

表 3、決策樹輸入格式

PID	輸入屬性																	目標值
	1896	1930	2076	2143	2149	2159	2259	2446	25	216	389	1499	1574	1574(1)	1613	1703	1703(1)	Seroconversion
L4007	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	Yes
L4048	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	Yes
L200	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	Yes
L3021	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	No
L208	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	No
L4056	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	No
L221	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	No
L239	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	No
L241	0	0	1	0	1	0	0	0	1	1	0	1	1	0	0	0	1	No
L2996	0	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0	1	No
L4035	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	No

接下來透過決策樹找出適合目標值的規則進行分類，以圖 5 為例，透過決策樹分析找出位置 1574 及位置 216 進行分類，其辨識正確率為 100%

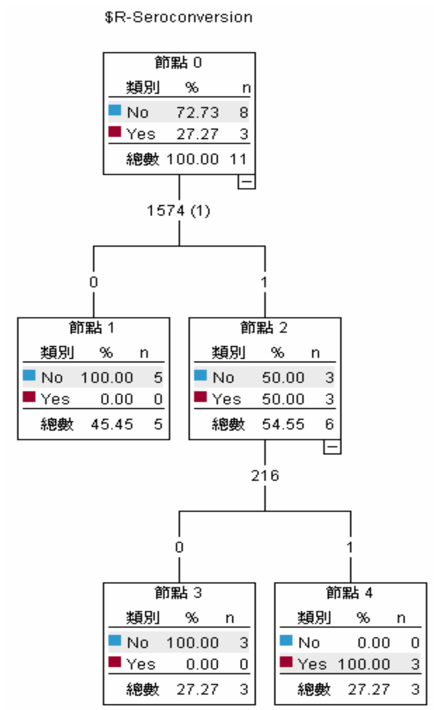

圖 5、決策樹分類結果

參、研究結果

一、依轉態進行規則分類

當 B 型肝炎同時產生 e 抗原 (HBeAg) 消失及抗體 (HBeAb) 出現則稱之為 e 抗原轉態 (HBeAg seroconversion)，在臨床上指的是 B 型肝炎病情已有好轉的現象。

本研究第一個部份，以 e 抗原是否轉態作為分類標準，產生決策樹如下圖 6 所示。

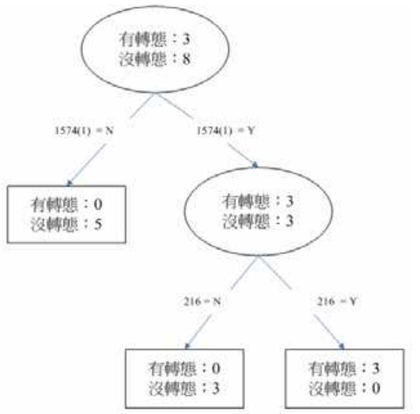

圖 6、轉態決策樹分類結果

由決策樹找出的分類結果，在 11 位病患中，不符合位置 1574 有 5 位病患皆為不轉態病患。符合特異點位置 1574 有 6 位病患，其中包括 3 位轉態及 3 位不轉態。從這 6 位病患再透過特異點位置 216 即可分類出符合此特異點的 3 位轉態病患及不符合特異點的另外 3 位不轉態。透過此兩個特異點即可將轉態及不轉態全部分類出來。接下來我們透過表 4 觀察決策樹所找出的特異點規則。

表 4、以轉態為分類標準之特異點規則分佈情況

特異點	抽血點鹼基表現				轉態			不轉態							
	baseline	VP1	AP1	BV2	L200	L4007	L4048	L208	L221	L239	L241	L2996	L3021	L4035	L4056
216	C	T	T	T	*	*	*	*			*	*			*
1574(1)	A	A	A	A	*	*	*		*	*				*	

從表 4 得知，特異點位置 216 其四個時間點的鹼基變化依序為 C、T、T、T，特異點位置 1574 其四個時間點的鹼基變化依序為 A、A、A、A，而轉態病患 L200、L4007、L4048，3 位皆符合此規則。另外，不轉態病患則在此兩個特異點上不會同時發生。因此透過此規則在本研究的第一部份其辨識正確率為 100%。

二、依基因型進行規則分類

B 型肝炎基因型態一共 8 類，台灣患者大多基因型 B、C 為主。本研究第二個部份以基因型 B、C 為分類標準，產生決策樹如下圖 7 所示。

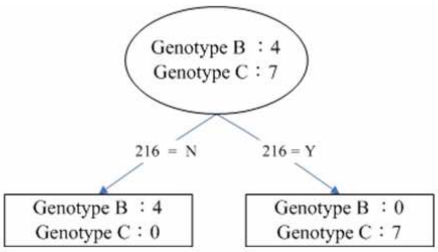

圖 7、基因型決策樹分類結果

由決策樹找出的分類結果，在 11 位病患中，屬於基因型 B 有 4 位病患必須符合特異點位置 216。若不符合特異點位置 216 時，則為基因型為 C 有 8 位病患。接下來我們透過表 5 觀察決策樹所找出的特異點規則。

表 5、以基因型為分類標準之特異點規則分佈情況

特異點	抽血點鹼基表現				基因型B				基因型C						
	baseline	VP1	AP1	BV2	L221	L239	L3021	L4035	L200	L241	L4007	L4048	L208	L2996	L4056
216	C	T	T	T					*	*	*	*	*	*	*

從表 5 得知，特異點位置 216 其四個時間點的鹼基變化依序為 C、T、T、T，而在基因型 C 的 7 名病患皆符合此一規則，反之在另外 4 名基因型 B 皆不符合此規則。因此透過一規則，在本研究的第二部份其辨識正確率亦為 100%。

三、依肝發炎指數上升倍數進行規則分類

肝發炎指數上升倍數值小於 5 倍為正常值，大於 5 倍為異常值，本研究第三個部份以肝發炎指數正常與否作為分類標準，產生決策樹如下圖 8 所示。

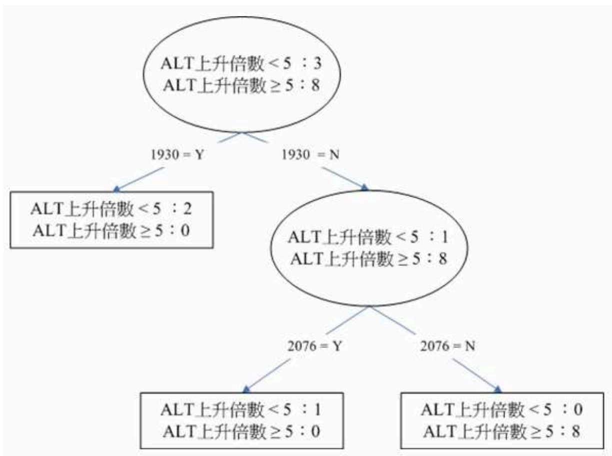

圖 8、肝發炎決策樹分類結果

由決策樹找出的分類結果，在 11 位病患中，符合特異點位置 1930 有 2 位病患皆為肝發炎指數正常病患。不符合位置 1930 有 9 位病患其中包括 1 位肝發炎指數正常病患及 8 位肝發炎指數異常病患。從這 9 位病患再透過特異點位置 2076 即可分類出符合此特異點的 1 位正常病患及不符合特異點的另外 8 位異常病患。透過此兩個特異點即可將肝發炎指數正常及肝發炎指數異常全部分類出來。接下來我們透過表 6 觀察決策樹所找出的特異點規則。

由決策樹找出的分類結果，在 11 位病患中，肝發炎上升指數正常的 3 位病患符合特異點位置 1930 或特異點位置 2076 其中之一。若不符合特異點位置 1930 及特異點位置 2076 時，則為肝發炎上升指數異常的病患有 8 位。

表 6、以肝發炎指數為分類標準之特異點規則分佈情況

特異點	抽血點鹼基表現				肝發炎指數正常			肝發炎指數異常							
	baseline	VP1	AP1	BV2	L221	L241	L2996	L200	L4007	L4048	L208	L239	L3021	L4035	L4056
1930	G	G	T	G	*		*								
2076	T	A	A	T		*	*								

從表 6 得知，特異點位置 1930 其四個時間點的鹼基變化依序為 G、G、T、G，特異點位置 2076 其四個時間點的鹼基變化依序為 T、A、A、T，而在決策樹找出的分類規則中，ALT 正常 3 位病患符合此兩個特異點之其中一種情形，而 ALT 異常的 8 名病患皆不符合這兩個特異點的突變狀況。因此透過此規則在本研究的第三部份其辨識正確率亦為 100%。

肆、討論與結論

針對本研究所找出的持異點位置(如表 7)進行討論。

e 抗原轉態部份特異點位置為 216 和 1574，位置 216 位於 S 基因片段，是表面抗原與人體免疫機制相關，它可能刺激免疫機制進行毒殺病毒，因此使病毒量降低、HBeAg 產量減少、免疫效率增加，HBeAb 產生因此達到 e 抗原轉態。而位置 1574 是 X 片段或 P 片段 RH 區域，X 片段上有許多控制 C 片段的活化區，例如：NER、C promoter 等。所以特異點位置 1574 維持穩定乃是病毒適應環境的演化趨向，但卻可能導致病毒易被免疫偵測達到轉態。而 e 抗原轉態與基因型 C 皆符合同一個特異點 216，跟過去研究中認為基因型 C 的 B 型肝炎病患較容易發生轉態相吻合。而肝發炎部份，特異點位置 1930 與 2076 都不影響病毒蛋白的變化。因此未能激發免疫反應，所以肝發炎表現正常。

表 7、B 型肝炎特徵與特異點關係及其辨識正確率

特徵	符合序列特異點	病患分類結果	辨識正確率
e 抗原轉態	216 & 1574	3/3	100%
基因型（C）	216	7/7	100%
ALT（正常）	1930 & 2076	3/3	100%

特異點序列表現分以下三個部份說明：

（一）以轉態而言

若病患符合序列位置 216 的四個抽血點，各時間點位置分別為 baseline 表現為 C、VP1 表現為 T、AP1 表現為 T、BV2 表現為 T 及位置 1574 的四個抽血點分別 baseline 表現為 A、VP1 表現為 A、AP1 表現為 A、BV2 表現為 A 時，則屬於轉態；若不符合則屬於不轉態。

（二）以基因型而言

若病患符合位置 216 的四個抽血點，各時間點位置分別為 baseline 表現為 C、VP1 表現為 T、AP1 表現為 T、BV2 表現為 T，則屬於基因型 C；若不符合位置 216，則屬於基因型 B。

（三）以肝發炎指數（ALT）而言

若病患符合位置 1930 的四個抽血點，各時間點位置分別為 baseline 表現為 G、VP1 表現為 G、AP1 表現為 T、BV2 表現為 G 及位置 2076 的四個抽血點，各時間點位置分別為 baseline 表現為 T、VP1 表現為 A、AP1 表現為 A、BV2 表現為 T 時，則屬於肝發炎正常；若不符合這兩個位置，則屬於肝發炎異常。

經本研究結果證實，確實可以從基因序列中找出時間變化的特異點與各 B 型肝炎特徵間的關係，其辨識正確率皆為 100%。在未來研究中除了可以陸續增加樣本數以求證本研究結果，也可以再增加其它問題進行分析，或許會有更重大的研究發現。

參考文獻

1. 王昶閔（民 95 年 5 月 17 日）。Pre-S 突變株 B 肝致癌兇手。自由時報，第 6 版。
2. 張玲霞、王永怡、饒文淑、韓俊英、雷周芸（民 84）。肝炎 300 問（初版）。臺北縣：世潮。
3. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Item in Large Database. ***Proceedings of the ACM-SIGMOD 1993 International Conference On the Management of Data, Washington, D.C., May 1993,*** 207-216.
4. Agrawal, R., & Srikant, R. (1994). Fast Algorithm for Mining Assoication Rule in Large Databases. ***Proceedings of the 20th International Conference On Very Large Database (VLDB) ,September 1994.*** 478-499.
5. Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. ***Proceedings of the International Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995,*** 3-14.
6. Chen, B.F. (2004). Clinical Significance of the Hepatitis B Virus Genotype. ***Fu Jen Journal of Medicine,*** 2: 273-281.
7. Frank, E. & Witten, H. (1999). ***Data Mining：Practical Machine Learning Tools and Techniques with Java Implementation .*** San Francisco：Morgan Kaufmann.
8. Han, J., & Kamber, M. (2001). ***Data Mining: Concepts and Techiniques.*** USA ： Morgan Kaufmann.
9. Karin, K.L., Yuzo, M., & Alistair H.K. (2002). Genetic variability in hepatitis B viruses. ***Journal of General Virology,*** 83: 1267-1280.