翁慈宗、劉冠良、韓昀達（2015），『多項式馬可夫簡易貝氏分類器結合狄氏先驗分配於基因序列分類之研究』，*中華民國資訊管理學報*，第二十二卷，第一期，頁 87-116。

多項式馬可夫簡易貝氏分類器結合狄氏先驗分配於基因序列分類之研究

翁慈宗
國立成功大學資訊管理研究所

劉冠良*
國立成功大學資訊管理研究所

韓昀達
國立成功大學資訊管理研究所

摘要

近年來隨著定序技術的發展，生物學家不再以傳統的方式進行生態環境的研究，而是由環境中擷取微生物的樣本，並且藉由定序技術瞭解物種的資訊，從中探索物種的多樣性。在 rRNA 序列分類的過程中，會利用 N-mer 的移動窗口對基因序列資料作特徵萃取，所萃取出的相鄰特徵會有 N-1 個字元重覆，因此萃取出的特徵集合具有關聯性，這與簡易貝氏分類器條件獨立的假設互相違背。本研究希望透過馬可夫簡易貝氏分器處理基因序列資料這種高維度與需要龐大運算需求的分類問題，不僅是因為馬可夫簡易貝氏分類器在運算效率上的優勢，也因為結合了馬可夫模型能夠改善簡易貝氏分類器在條件獨立假設的問題，其中本研究採用多項式模型作為機率模型，在概似機率的計算上考慮了特徵頻率，而會有較佳的分類表現。此外，本研究加入了先驗分配－狄氏分配，期望藉由馬可夫簡易貝氏分類器和狄氏分配的結合，透過兩種先驗分配參數－分子先驗分配參數與分母先驗分配參數的設定，提升分類正確率。

本研究以兩種不同的方式－狄氏分配_分子分母與狄氏分配_分母分子對四個基因序列資料檔來作測試。實證結果發現，狄氏分配_分子分母，在同一個類別值內先進行分子參數的調整，再進行分母參數的調整會有較好的分類結果。該兩種方法在參數調整完畢後，其分類正確率已高於 RDP 分類器，相較於簡易貝氏分類器結合狄氏分配，多了分母先驗分配參數可供調整，因此有較高的分類結果。

關鍵詞：狄氏分配、馬可夫模型、簡易貝氏分類器、核甘酸基因序列

* 本文通訊作者。電子郵件信箱：kliangliu@gmail.com
2010/03/13 投稿；2014/07/07 修訂；2014/10/01 接受

Wong, T.T., Liu, K.L. and Han, Y.D. (2015), 'Dirichlet Priors for Markov Naïve Bayesian Classifiers with Multinomial Model for Gene Sequence Data', *Journal of Information Management*, Vol. 22, No. 1, pp. 87-116.

Dirichlet Priors for Markov Naïve Bayesian Classifiers with Multinomial Model for Gene Sequence Data

Tzu-Tsung Wong
Institute of Information Management, National Cheng Kung University

Kuan-Liang Liu*
Institute of Information Management, National Cheng Kung University

Yun-Da Han
Institute of Information Management, National Cheng Kung University

Abstract

Purpose－The RDP classifier is computationally efficient and does not require sequence alignment. It also works well with short sequences and provides a unique niche for applications using the NGS technologies that generate millions of short sequences. The performance however, is hampered by the conditional independent assumption on features. The dependency is especially obvious in attributes by which the k-mer method extracts sequences from sequences with sliding window where each attribute is overlapped by k-1 base with its previous and next attribute.

Design/methodology/approach－This study developed a multinomial Markov-based Bayesian classifier which remedies the unrealistic independent assumption by Markov model. In order to prevent probability estimate of feature to become zero and distort the classification result, Laplace estimate is usually utilized as a prior for all features. However, the setting assumes a fix confidence level for all features. In this study, we further develop a noninformative generalized distribution for prior setting that will allow different confidence level settings for different features.

* Corresponding author. Email: kliangliu@gmail.com
2010/03/13 received; 2014/07/07 revised; 2014/10/01 accepted

Findings－The experimental results on bacterial 16S and fungal 28S rRNA gene sequence sets show that the proposed model can achieve higher prediction accuracy than the well-known RDP classifier in all ranks. Since the number of priors for a class value in the Markov naïve Bayesian classifier is two instead of one in the naïve Bayesian classifier, the best noninformative Dirichlet priors do enhance the performance of the Markov naïve Bayesian classifier.

Research limitations/implications － The study proposes to model DNA sequences as a k^{th} order Markov chain on the alphabet A,C,G and T. That is, the probability of observing a particular symbol only depends on the k previous one. Since under this model, the probability of a read can be written as a ratio of products of the probability of overlapping k-mers, it does introduce additional computational overheads to the current implementation of the RDP classifier. However, the overhead is just one time calculation during training process practically.

Practical implications－Since the ability to obtain thousands of rRNA sequences from environmental and Human Microbiome Project samples using high-throughput sequencing technologies has become a reality, accurate sequence classification is a critical component of ecological interpretation of environmental datasets. The approach used in this article to evaluate bacterial and fungal sequences proves to be a valuable tool to determine the most important factors affecting classification accuracy.

Originality/value － This study develops a Markov Bayesian classifier that remedies the strong conditional assumption on naïve Bayesian classifier by Markov assumption. Since the k-mer method uses an overlapping sliding window to extract features from raw sequence data, the conditional independent assumption on features is clearly violated. In the opposite, the Markov model that assumes each feature is dependent on previous features fits k-mer extraction method perfectly.

Keywords: Dirichlet distribution, Markov model, naïve Bayesian classifier, rRNA gene sequence

壹、緒論

近年來隨著新世代定序技術（next-gen sequencing; NGS）的發展，科學家不再以傳統的方法在實驗室中培養單一物種（Handelsman et al. 1998），而是從生態環境中直接取得樣本，使得科學家能更進一步探索微生物。在基因序列的分類問題上，經常使用 N-mer 方法以移動窗口來萃取特徵，而由於基因序列是由 4 種核苷酸所組成，若 N 設定為 8，資料矩陣維度將是 65536，在加上 rRNA 基因序列長度介於 1500bp~2000bp，產生的資料矩陣不僅維度高而且稀疏（sparse）。簡易貝氏分類器（naïve Bayesian classifier）因其計算效率且容易運用，常被廣泛使用在此分類領域上，Wang 等（2007）即利用簡易貝氏分類器建立了 16S rRNA 的 RDP 分類器（Ribosomal Database Project; RDP），其分類迅速且分類正確率較傳統的序列比對工具 BLASTN（Camacho et al. 2009）高，因此近幾年來被廣泛使用於各種人類微生物組計劃（human microbiome project; HMP）（Human Microbiome Project Consortium 2012）以及地球微生物組計劃中（earth microbiome project; EMP）（Gilbert & Meyer 2012）等定序計劃中，而建立在相同的分類模型上，Liu 等（2012）也證明了簡易貝氏分類器在真菌 28S rRNA 序列資料集的適用性。

但是簡易貝氏分類器在基因序列的分類上，會面臨萃取出的特徵集合是否具有關聯性以及同一序列特徵重複出現的兩種問題。

1. 簡易貝氏分類器假設所萃取出的特徵集合間彼此獨立，但是萃取序列特徵所使用的移動窗口是以 N-mer 個字元為長度進行切割，切割出來的兩兩特徵彼此有(N-1)-mer 的字元相同，會與簡易貝氏分類器的條件獨立假設互相違背。
2. 在基因序列的特徵萃取上，通常所使用的機率模型只考慮到該特徵是否有出現過，而不考慮該特徵所出現的次數。

針對第一個問題，本研究採用馬可夫假設來處理特徵集合之間為非獨立的狀況，馬可夫模型對於事件的相依假設和序列萃取方式一致；而針對第二個問題，Liu 與 Wong（2013）利用多項式模型，考量了特徵出現頻率的觀念，解決同序列特徵重複的問題，而該研究中所獲得的分類正確率比經常使用在分析基因序列的 RDP 分類器（ribosomal database project; RDP）來得高，但是該分類器是使用簡易貝氏分類器，因此有條件獨立假設在特徵萃取上有較不適合的問題。因此如何將上面的技術作結合，解決基因序列關聯性問題與同序列重複特徵出現之問題是值得去探討的。

特徵萃取是使用移動窗口來切割特徵，切割出的特徵集合彼此具有關聯性，與簡易貝氏分類器的條件獨立假設會有違背，根據此問題，本研究採用馬可夫假

設來處理此問題；而在基因序列的特徵萃取上，常採用二元模型，然二元模型只考慮特徵出現與否，忽略該特徵出現的次數，而多項式模型考量了特徵出現的次數，根據 McCallum 等（1998）指出，多項式模型相較有較好的分類正確率，且 Liu 與 Wong（2013）亦證明了多項式模型運用於簡易貝氏分類器可以在 rRNA 分類問題上獲得較高的分類正確率。此外本研究加入先驗分配，如狄氏分配（Dirichlet distribution），藉由設定相關參數的方式，估計出各個可能值的先驗機率，來解決基因序列因特徵數量眾多，在某些類別值下，許多特徵未曾出現，造成計算條件機率時因為連乘的結果，會有推估該類別的事後機率值為 0 的問題。

貳、文獻探討

一、多源基因體學

多源基因體學的發展主要歸因於近年來各種新世代定序技術（next-gen sequencing; NGS）的突破，相較於以往在實驗室培育單一物種的分析方式，研究人員現在能夠以低成本大量的對特定微生物聚落進行定序，透過這樣的方式，研究人員可以針對特定微生物群聚進行不同時間、不同地點的大規模分析，探索微生物群聚與地球上各種物種的共生關係。

定序過程中會產生大量的序列，如何有效的將這些序列分類到正確的微生物種，成為以多源基因體分析微生物群聚相當重要的步驟，Wang 等（2007）以簡易貝氏分類器作為基礎，發展了 RDP 分類器，其具有分類精準且快速的優勢，因此常作為基因序列的分類工具。而 Rosen 等（2008）提出了 N-mer frequency，結合了簡易貝氏分類器作為分類工具，N-mer frequency 以移動窗口的方式，機率推估的方式為紀錄特徵出現的頻率。隨後 Rosen 等（2011）更以簡易貝氏分類器作為網路應用服務，幫助使用者進行數據分析。

由於 RDP 分類器不管在分類正確率或是分類速度上都表現的比傳統的序列比對工具 BLASTN 優異（Liu et al. 2012），因此近幾年來被廣泛使用於各種人類微生物組計劃（human microbiome project; HMP; http://commonfund.nih.gov/hmp/index）（Human Microbiome Project Consortium 2012）以及地球微生物組計劃（earth microbiome project; EMP; http://www.earthmicrobiome.org/）（Gilbert et al. 2012）等各種微生物定序計劃中，其中人類微生物組計劃是一個由美國國家衛生研究院（national institute of health; NIH）所主導的研究計劃，其透過新世代定序技術進行人類微生物組的定序工作，研究腸道、口腔、皮膚、鼻腔、生殖道中和人體共生的微生物，這項計畫在科學史上，是繼人類基因體計畫的另一個里程碑；而相較於人類基因體計劃研究微生物與人體的互生關係，地球微生物組計劃則是對各種環境做定序，包含海洋、土壤、沙漠、雨林甚至極地等，藉此了解微生物在這些

環境中所扮演的關鍵角色。

二、核糖體 RNA（ribosomal RNA, rRNA）

自從 Woese 與 Fox（1977）藉由 16S rRNA 基因序列所建立的演化樹，發現生命樹的新分枝－古菌後，rRNA 的相關特性與分析便開始受到重視，並漸漸的成為近代分子生物學中分析微生物的重要依據。此分子序列之所以擁有區分不同物種的高鑑別力，在於其擁有下列特性（Woese 1987）：

- 核糖體細胞內最多的胞器，在原核及真核生物鐘皆可發現，其所負責合成蛋白質屬於生物體相當基本且重要的功能，因此廣泛分佈於所有生命個體中。
- rRNA 基因序列中的保守區域（conserved regions）可用於構建所有生命演化樹，而變異區域（variable regions）可用來區別屬或者種。
- rRNA 在生物體的功能相當基本且重要，需要精細調控的轉錄機制才能夠正常實現，因此基因的水平轉移（horizontal transfer）非常難發生（Olsen et al. 1986）。

細菌的 rRNA 中，16S rRNA 長度適中，基因片段保守性高，Woese（1987）稱其為生物演化的分子時鐘（molecular chronometer），可以提供微生物在演化與鑑別上足夠的訊息，因此近幾年 16S rRNA 基因序列仍然是分子生物學領域最普遍被用來分析微生物的依據。

早期的物種鑑別主要是根據生物外觀來判定，然而在科學家開始從分子生物學的角度分析時，卻發現許多以往被叛定為同一物種的生物，在序列比對上卻十分岐異，因此目前有許多科學家期望能用更精細與嚴謹的方法，利用已知的文獻與序列比較資訊為基礎，重新定義分類架構，這樣的過程稱為 curation。而目前較知名的 curator 包括有 NCBI（National Center for Biotechnology Information）、RDP（http://rdp.cme.msu.edu/index.jsp）（Cole et al. 2014）、the ARB project（http://www.arb-home.de/）（Ludwig et al. 2004）、以及 the Greengenes project（http://greengenes.lbl.gov）（DeSantis et al. 2006）。

與細菌 16S rRNA 類似，真菌 28S rRNA 近幾年來也開始被使用來在相關研究中鑒別真菌物種，而因應這樣的需求，Liu 等（2012）建立了第一個真菌 28S rRNA 公開資料集，並且利用 RDP 分類器測試該資料集在不同長度，不同引子位置的分類正確率，該資料集目前也已提供給 RDP 作為公開測試訓練資料，任何研究人員都可以利用該資料集以 RDP 分類器對其所定序的資料進行分類。本研究將使用 RDP 所提供的 16S rRNA 與 28S rRNA 序列資料。

三、簡易貝氏分類器

簡易貝氏分類器是貝氏理論加上條件獨立的假設，具有速度快和面對龐大運算量有良好的效能，特別適合使用在維度高的分類問題上。

假設欲分類的一筆資料 x 具有 p 個特徵，表示式為 $x=(x_1,x_2,...,x_p)$，而資料 x 的類別值有 n 種可能，類別值集合為 $\{c_1,c_2,...c_n\}$，其中 c_n 代表第 n 個類別可能值，在給一筆定資料 x 下，若我們欲推估資料 x 屬於何種類別值，則可以帶入以下貝氏分類器的計算公式

$$P(c_j \mid x)=\frac{p(c_j,x)}{p(x)}=\frac{p(x \mid c_j)p(c_j)}{p(x)} \tag{1}$$

在式子(1)的 $p(c_j \mid x)$ 為事後機率（posterior probability），表示給定資料 x 的條件下，屬於類別值 c_j 的機率，$p(x \mid c_j)$ 為概似機率（likelihood probability），則 $p(c_j)$ 稱為事前機率（prior probability）。因為都為同一筆資料 x，所以 $p(x)$ 是固定的，可以將 $p(x)$ 視為常數，將(1)的式子簡化為

$$p(c_j \mid x) \propto p(x \mid c_j)p(c_j) \tag{2}$$

其中所獲得的事後機率值滿足以下式子

$$c^*=\alpha r\ g\max_{cj}\ p(c_j \mid x) \tag{3}$$

在各個所計算出的事後機率中挑選出最大的機率值，而該類別值 c_j 會指定給 c^*，則資料 x 的類別值就會被預測為 c_j 這個類別。

而貝氏分類器加上了條件獨立的假設即為簡易貝氏分類器。條件獨立的假設代表的是在給定某類別值 c_j 下，各個特徵為互相獨立，即為特徵值 x_i 不影響特徵值 x_i，因此根據條件獨立的假設可以將式子(3)展開後可以得到事後機率

$$p(c_j \mid x) \propto p(x_1 \mid c_j)\times p(x_2 \mid c_j)\times ...\times p(x_n \mid c_j)\times p(c_j)=\prod_{i=1}^{n} p(x_i \mid c_j)p(c_j) \tag{4}$$

其中獲得的事後機率值滿足以下式子

$$c^*=\alpha r\ g\max_{cj}\ p(c_j)\prod_{i=1}^{n} p(x \mid c_j) \tag{5}$$

因此只要計算得出各個 $p(x \mid c_j)$ 和 $p(c_j)$ 的機率值，就可以推估樣本資料 x 應屬於

哪種類別。

依據 Chen 等（2009）的文獻，在文件分類上經常採用的主要有兩種機率模型：二元模型（binomial models）和多項式模型（multinomial models），這兩種模型最大的差別在於計算概似機率的不同，二元模型為考慮字彙是否有出現，多項式模型是計算字彙的出現頻率。而文件分類與基因序列資料的分類因為在概念與做法上類似，接下來分別為二元模型和多項式模型作介紹。

1. 二元模型

在二元模型中，探討字彙的方式為是否出現該字彙，若文件中出現該字彙，則出現次數 $f_i = 1$，相反若沒有出現該字彙，則 $f_i = 0$，即 $f_i \in \{0,1\}$。

因為概似機率計算方式的不同會由式子(5)修改為

$$\begin{aligned} c^* &= \alpha r\ g \max_{cj}\ p(c_j)\prod_{i=1}^{n} p(w \mid c_j) \\ &= \alpha r\ g \max_{cj}\ p(c_j)\prod_{i=1}^{n} p(w_i \mid c_j)^{fi}[1-p(w_i \mid c_j)]^{(1-fi)} \end{aligned} \tag{6}$$

其中 $p(w_i \mid c_j)$ 可以由之前的訓練資料估算而得。

2. 多項式模型

在多項式模型中，探討字彙的方式為字彙的出現頻率，字彙出現字數 $f_i \in \{0,1,2...\}$。多項式模型的概似機率會將式子(5)表示為

$$\begin{aligned} c^* &= \alpha r\ g \max_{cj}\ p(c_j)\prod_{i=1}^{n} p(w \mid c_j) \\ &= \alpha r\ g \max_{cj}\ p(c_j)(\sum_{i=1}^{n} x_i)!\prod_{i=1} \frac{P(w_i \mid x)^{fi}}{x_i!} \end{aligned} \tag{7}$$

其中 $(\sum_{i=1}^{n} x_i)$ 為固定值可以視為常數，最後式子(7)簡化為下面式子

$$c^* = \alpha r\ g \max_{cj}\ p(c_j)\prod_{i=1}^{n} p(w_i \mid c_j) \tag{8}$$

其中機率 $p(w_i \mid c_j)$ 可由之前訓練資料計算而得。

Wang 等（2007）的 RDP 分類器與 Rosen 等（2008）的 Rosen 分類器，兩者都是以簡易貝氏分類器為基礎的分類方式，兩者的差異在於前者是使用二元模型，後者使用的是多項式模型。

四、馬可夫簡易貝氏分類器

馬可夫模型是一種利用過去已知的資料，來對未來作預估結果的方法，目前

在統計學、地理學、音訊辨識與生物基因上都被積極廣泛地使用（Baum and Eagon 1967; Brady et al. 2009; Kotamarti et al. 2010; Ghosh et al. 2012），對未來的結果作預測與推估。

本研究將馬可夫模型結合至簡易貝氏分類器，對原本條件獨立假設的部分進行修正，改善簡易貝氏分類器條件獨立假設的問題。以下將為馬可夫模型和馬可夫簡易貝氏分類器的基本運作流程做介紹。

假設有一筆基因序列資料的一個字元 s_m ，s_m 的機率為 $p(s_m \mid s_1, s_2, ..., s_{m-1})$，在推估 s_m 時需要前面 m-1 個字元來推估，因此每推估一個字元就需要前面所有字元的資料才能推估下一個字元的機率。而馬可夫模型能夠將這樣的推估方式進行簡化，只需要要前面 k 個字元就可以推估該字元的轉移機率，因此可以將 s_m 的機率表示成以下形式

$$p(s_m \mid s_1, s_2, ..., s_{m-1}, c_j) = p(s_m \mid s_{m-k}, s_{m-k+1}, ..., s_{m-1}, c_j) \tag{9}$$

同樣地可以把馬可夫模型的概念用來計算特徵的條件機率，特徵 $w_m^{(N)}$ 代表的是長度為 N-mer 的第 m 個特徵，而透過馬可夫模型的概念，特徵 $w_m^{(N)}$ 的推估只需考慮前一個特徵 $w_{m-1}^{(N)}$ ，而不需要將先前全部的特徵一併可慮，因此 $w_m^{(N)}$ 的轉移機率修改成以下形式

$$p(w_m^{(N)} \mid w_1^{(N)}, w_2^{(N)}, ..., w_{m-1}^{(N)}, c_j) = p(w_m^{(N)} \mid w_{m-1}^{(N)}, c_j) \tag{10}$$

因此一筆基因序列資料 r 的機率透過馬可夫模型可以表示為

$$\begin{aligned} p(r \mid c_j) &= p(w_1^{(N)}, w_2^{(N)}, ..., w_m^{(N)}, c_j) \\ &= p(w_1^{(N)} \mid c_j) p(w_2^{(N)} \mid w_1^{(N)}, c_j) p(w_3^{(N)}, c_j) ... p(w_m^{(N)} \mid w_{m-1}^{(N)}, c_j) \end{aligned} \tag{11}$$

又因為

$$p(w_m^{(N)} \mid w_{m-1}^{(N)}, c_j) = \frac{p(w_m^{(N)}, w_{m-1}^{(N)} \mid c_j)}{p(w_{m-1}^{(N)} \mid c_j)} \tag{12}$$

而分子的部分，因為特徵 $w_m^{(N)}$ 和特徵 $w_{m-1}^{(N)}$ 所包含的字元剛好可以結合表示為特徵 $w_m^{(N+1)}$ ，因此分子的部分表示為以下型式

$$p(w_m^{(N)}, w_{m-1}^{(N)} \mid c_j) = p(w_m^{(N+1)} \mid c_j)$$

因此式子(12)可以修改成

$$p(w_m^{(N)} \mid w_{m-1}^{(N)}, c_j) = \frac{p(w_{m-1}^{(N+1)} \mid c_j)}{p(w_1^{(N)} \mid c_j)} \tag{13}$$

最後將式子(13)帶回式子(11)得到以下公式

$$p(r \mid c_j) = p(w_1^{(N+1)} \mid c_j) \frac{p(w_2^{(N+1)} \mid c_j)}{p(w_2^{(N)} \mid c_j)} \ldots \frac{p(w_m^{(N+1)} \mid c_j)}{p(w_{m-1}^{(N)} \mid c_j)} \tag{14}$$

因此只要把萃取出的第一個特徵 $p(w_1^{(N)} \mid c_j)$ 的機率與各個 $\frac{p(w_m^{(N+1)} \mid c_j)}{p(w_m^{(N)}, c_j)}$ 的機率計算得出，就可以推估出該筆序列是屬於何種物種。

五、狄氏分配

狄氏分配為定義在單位體（unit simplex）上的多變量分配（multivariate distribution），目的是藉由先驗分配參數的推估，來使分類正確率能夠有效的提升。先驗分配是一種以知識性或以先前經驗的方式來幫助、改善所欲推估之估計值。而單位體的意思代表的是各個變數皆不可為負值，而且總和必須為 1。狄氏分配由於一般動差函數（general moment function）的關係，讓狄氏分配的運算複雜度降低，更容易去推估機率值，因此經常作為先驗分配使用於簡易貝氏分類器中。

若有一特徵具有 k 個可能值，則各個可能值的機率值集合 θ 可以表示為 $\{\theta_1, \theta_2, ..., \theta_k\}$，其中 θ 滿足 $\theta_1 + \theta_2 + \theta_3 + ... + \theta_k \le 1$ 與 $\theta_i > 0 (i = 1, 2, ..., k)$。藉由機率密度函數（probability density function）可以推估 θ 的分佈狀況，並且服從狄氏分配，表示為 $\theta \sim GD_k(a_1, a_2, ..., a_k; a_{k+1})$，則其機率密度函數表示式如下；

$$f(\theta) = \frac{\Gamma(a^*)}{\prod_{j=1}^{k+1} \Gamma(a_j)} \prod_{j=1}^{k} \theta_j^{a_{j-1}} (1 - \theta_1 - \theta_2 - ... - \theta_j)^{a_{k+1}-1} \tag{15}$$

其中 $a^* = a_1 + a_2 + ... + a_{k+1}$，$a^*$ 為各個參數的總和。

而根據（Wong 1998）的動差函數可以將 θ 的期望值表示成以下式子：

$$E(\theta_i) = \frac{a_i}{a^*}, i = 1, 2, ..., k \tag{16}$$

$$E(\theta_{k+1}) = \frac{a_{k+1}}{a*} \tag{17}$$

透過學習資料 O 計算出各個特徵可能值的出現次數，會將狄氏分配進行修正，a_i 重新計算，以下為 a_i 的計算式子

$$a_i' = a_i + ct_i \tag{18}$$

最後 θ_i 的期望值的計算式子如下：

$$E(\theta_i \mid O) = \frac{a_i + ct_i}{a* + allct}, i = 1,2,...,k+1 \tag{19}$$

其中 *allct* 代表的是某特徵內，所有特徵值出現次數的總合，而 $p(\theta \mid O)$ 會服從修正後的狄氏分配，並表示為 $\theta \mid O \sim D_k \theta \mid O \sim GD_k(a_1', a_2', ..., a_k'; a_k')$。

而狄氏分配所使用的條件也具有限制，Wong（1998）推導一般動差函數，證明出任兩個隨機變數彼此間具有負相關的特性，即負相關需求（negative-correlation requirement）。Wong（2009）提到等信賴原則（equal-confidence requirement），當狄氏分配作為先驗分配時，利用正規化變異數（normalized variance, NV）衡量各變數的信心水準（confidence level），而各變數的正規化變異數必須相同，使得所有變數的信心水準也需相同，因此在調整信心水準時，狄氏分配調整的是整體的信心水準，無法依據各變數做單獨的調整，所以在使用狄氏分配時會有負相關需求與等信賴原則的限制條件。

參、研究方法

一、資料前置處理

本研究使用特徵萃取的方式為 N-mer frequency，假設具有 j 筆基因序列資料，表示成 $S = \{S_1, S_2, ..., S_j\}$，萃取出所有的特徵集合 F，表示成 $F = \{F_1, F_2, ..., F_m\}$，其中 m 代表的是特徵集合中具有 m 個特徵，而 m 會等於 4^N，因此移動窗口的長度 N 越大，可能的特徵集合數量 m 也越大。

二、多項式馬可夫簡易貝氏分類器

在本研究流程中的學習階段與測試階段上，皆採用以多項式模型為基礎的馬可夫簡易貝氏分類器，除了簡易貝氏分類器在運算效率上十分適合處理高維度的分類問題之外，相較於 Wang 等（2007）所提出的以二元模型作為基礎的簡易貝氏

分類器，本研究將採用以多項式模型作為分類器的機率模型，原因在於多項式模型相較於二元模型在概似機率的計算上更加考慮了特徵的出現頻率，而在分類結果上會有較好的表現，因此本研究採用以多項式模型為基礎的馬可夫簡易貝氏分類器作探討。

根據先前馬可夫簡易貝氏分類器的式子(14)，可以將每個特徵的機率值，透過多項式模型表示為式子(20)：

$$p(r \mid c_j) = p(w_1^{(N+1)} \mid c_j) \frac{p(w_2^{(N+1)} \mid c_j)}{p(w_2^{(N)} \mid c_j)} \cdots \frac{p(w_m^{(N+1)} \mid c_j)}{p(w_{m-1}^{(N)} \mid c_j)}$$

$$p(w_j^{(N)} \mid c_j) = \frac{f(w_j^{(N)} \mid c_i)}{v(c_i)} \tag{20}$$

由於本研究採用馬可夫簡易貝氏分類器，因此在式子(14)中可以發現分子與分母的特徵長度並不相同，位於分子的特徵 $w_j^{(N+1)}$ 為(N+1)-mer 的特徵，而位於分母的特徵 $w_j^{(N)}$ 則是為長度 N-mer 的特徵，這也是在往後進行先驗分配參數的調整時，將會有兩個先驗分配參數需要調整。

而透過多項式模型，可以將式子(14)中，每個特徵的機率值 $p(w_j^{(N)})$ 表示為式子(20)，其中 $f(w_j^{(N)} \mid c_j)$ 代表的是在類別值 c_i 下 $w_j^{(N)}$ 所出現的次數，而 $v(c_i)$ 表示為類別值 c_i 所擁有全部特徵出現的總和，若類別值 c_i 具有 16 筆序列資料，而每一筆序列資料所出現的特徵次數都各出現 100 次，因此 $v(c_i)$ 的特徵出現次數總合為 $16 \times 100 = 1600$，其中在採用狄氏分配作為先驗分配時，會將先驗分配參數 a_c 加入式子(20)中進行參數的調整，以尋找最佳的參數。

其中值得注意的是多項式模型與二元模型的不同就在於 $f(w_j^{(N)} \mid c_j)$ 條件機率的計算，多項式模型的特徵出現次數 $f(w_j^{(N)} \mid c_j) \in \{0,1,2,...\}$，而二元模型的特徵出現次數 $f(w_j^{(N)} \mid c_j) \in \{0,1\}$，本研究將採用多項式模型針對特徵出現頻率 $f(w_j^{(N)} \mid c_j)$ 進行計算。

三、狄氏分配參數的調整方法

在本研究中，以狄氏分配作為先驗分配時，先驗分配參數調整的方式是先對各個類別值進行排序的動作，依影響力的大小對各個類別值進行排序，影響力越大者則排序在越前面，隨後再依據類別值的順序依序進行參數上的調整。而類別值排序的方式則是根據該類別所擁有的序列數量而定，擁有序列數量越多的類別值代表該類別越為重要，影響力程度也越大。

而使用狄氏分配時，由於不同類別值下的特徵，其條件機率下的類別值並不相同，又因為本研究使用馬可夫簡易貝氏分配器，根據式子(14)在計算特徵的機率時會有分母與分子在特徵長度上的不一致，分母為長度 7-mer 的特徵，分子為長度 8-mer 的特徵，因此在同一個類別值中所需調整的先驗分配會具有兩個，其一為分母的先驗分配，另一個為分子的先驗分配，因此先驗分配的數量為類別值個數的兩倍。

而加入狄氏分配作為先驗分配時，先驗分配參數 a_c 將加入到式子(20)中並表示成以下式子：

$$p(w_j^{(N)} \mid c_j) = \frac{f(w_j^{(N)} \mid c_j) + a_c}{v(c_j) + 4^N a_c} \tag{21}$$

其中 4^N 代表的是所有可能特徵集合的數量，若是以長度為 8-mer 的移動窗口進行特徵萃取，則萃取出可能特徵集合的數量為 $4^8 = 65536$；而本研究由於使用馬可夫簡易貝氏分類器，因此分子與分母所有可能的特徵集合數量並不相同，分子的特徵長度為 8-mer，則所有可能特徵集合的數量為 $4^8 = 65536$；而分母的特徵長度為 7-mer，因此所有可能特徵集合的數量為 $4^7 = 16384$。

一開始，將各個類別值的參數 a_c 都設為 0.000001，先驗分配參數調整的順序以類別值排序後的順序開始依序調整。首先第一個類別值會有分母的先驗分配參數與分子先驗分配的參數，本研究在分母與分子先驗分配參數的調整順序會分別採用以下兩種方式進行測試，第一種是先進行分子部分的調整，再進行分母部分的調整；第二種是先調整分母的先驗分配參數，再調整分子的部分，而該兩種方式會在往後的研究中進行測試。在決定完先行調整分子或分母的先驗分配參數之後，該參數 a_c 會從 0.01 開始以每次增加 0.01 的方式進行直到 a_c 等於 1.000001 即停止，而每次增加 0.01 後會隨即進行一次 leave one out 並將分類正確率紀錄起來，取過程中最好分類正確率時的參數 a_c 為最佳設定參數，並將此 a_c 值固定住，接著調整第一個類別值中另一個先驗分配參數，也是以 a_c 值從 0.01 至 1 的過程中，取其最佳分類正確率的 a_c 值為最佳參數，並且固定住。在第一個類別值的兩個先驗分配參數都設定完畢後，會依照此規則尋找下一個類別值的先驗分配參數，直到全部類別值的先驗分配參數都已設定完成。

四、驗證方式

本研究將採用 leave one out 的驗證方式，在每一次驗證的過程中會去除一筆序列資料當作測試資料集，而其餘的（N-1）筆序列資料作為訓練資料集進行學習，

再對該筆基因筆序列資料進行分類。最後將所有的序列資料依序進行驗證，並統計出所有序列資料的成功次數，由於每筆基因序列資料都進行一次驗證，因此除以資料檔中全部基因序列資料的筆數，即可得到在某先驗分配參數組合下的分類正確率。

肆、實證研究

一、rRNA 測試資料檔

本研究所使用的資料檔為細菌和真菌 rRNA 資料檔，序列筆數分別為 4354 筆和 8506 筆，細菌 16S rRNA 序列長度大約是 1600-bp，真菌 28S rRNA 序列長度則大約是 1400-bp，兩筆序列資料檔自 RDP 的 sourceforge 頁面（http://sourceforge.net/projects/rdp-classifier/files/RDP_Classifier_TrainingData/）中下載。在資料進行前置處理時，將只具有一筆基因序列資料的類別值 singleton 去除不予計算。Bacteria 資料檔經過資料整理後，具有 3672 筆基因序列資料，包含 500 個類別值；Fungi 資料檔經過整理後，樣本數共有 7730 筆基因序列資料，包含 926 個類別值，往後以 Bacteria3672 和 Fungi7730 稱之。此外，本研究更將該兩個資料檔整理為序列筆數大於 10 筆的類別資料，降低先驗分配調整的調整個數並與原始資料筆數的資料檔進行比較，篩選後的 Bacteria 樣本數為 2035 筆序列資料，包含 79 個類別值；而 Fungi 資料檔經過篩選後，共具有 4954 筆序列資料，含有 159 個類別值，該兩個資料檔以 Bacteria2035 和 Fungi4954 稱之。

二、馬可夫簡易貝氏分類器之實證研究

本研究在馬可夫簡易貝氏分類器中加入了 Laplace Estimate，避免某個類別值的事後機率值 0 的情況發生，而在初始先驗分配參數的設定上，分子先驗分配參數與分母先驗分配參數都設定為 0.000001，讓相同的先驗分配參數具有一樣的信心水準，並於馬可夫簡易貝氏分類器中將以五種不同的參數值 0.01, 0.001, 0.0001, 0.00001, 0.000001 分別依序作測試，以下表 1 為資料檔 Bacteria 和 Fungi 的成功次數與分類正確率比較表。

從表 1 的數據資料中，可以發現為類別值序列筆數大於 10 筆的資料檔 Bacteria2035 和 Fungi4954 相較於原本 Bacteria3672 與 Fungi7730 的分類正確率來的高。我們認為正確率提昇可能的原因有兩個，首先，資料檔中剔除了小於 10 筆基因序列資料的類別後，會使在訓練資料集中每個類別值都擁有大於 10 筆的訓練資料筆數，因此在測試進行計算條件機率時，會有較好的分類正確率；另一方面，由於資料經過篩檢後，細菌與真菌資料集的類別個數分別由 500 和 926 降到 79 和

159，資料類別數的降低使得整體資料複雜度也降低，因此也有可能獲得較好的分類正確率。

另外，由比較表中更可以發現不同先驗分配的參數值會有不同的分類正確率，其中從數據資料可以觀察出，各個 Bacteria 和 Fungi 資料檔參數設定為 0.000001 時，會比先驗分配參數為 0.01 時，有較好的分類正確率。因此由此可知在使用馬可夫簡易貝氏分類器時，參數值為 0.01 下所擁有的信心水準相對於參數值為 0.000001 時來得太高，也驗證了先驗分配參數的調整會有較好的分類正確率。

馬可夫簡易貝氏分類器於 Bacteria 和 Fungi 資料檔中，在不同參數值下的分類正確率相較於 RDP 分類器都來得低，其中可能的原因為馬可夫簡易貝氏分類器結合了馬可夫模型並且加上了多項式機率模型，在機率值的估計上，相較於簡易貝氏分類器在在機率值的估算上，馬可夫簡易貝氏分類器多除以 7-mer 分母的機率值，進而可能影響了原先 8-mer 分子的機率值，而使最後的分類正確率降低。然而馬可夫簡易貝氏分類器有分子先驗分配參數與分母先驗分配參數可進行調整，參數調整的彈性較簡易貝氏分類器來的高，因此下一小節將以馬可夫簡易貝氏分類器結合狄氏分配，以調整類別值的分子與分母先驗分配參數，來提升分類正確率。

表 1：各個資料檔在五種不同參數值下與 RDP 分類器的分類正確率比較

資料檔	馬可夫簡易貝氏分類器					RDP 分類器
	參數 0.01	參數 0.001	參數 0.0001	參數 0.00001	參數 0.000001	
Bacteria2035	93.61%	94.20%	95.08%	95.52%	95.77%	97.29%
Bacteria3672	84.23%	86.95%	88.56%	89.46%	90.08%	93.65%
Fungi4954	86.71%	86.57%	87.18%	87.60%	87.90%	88.02%
Fungi7730	70.53%	74.06%	74.42%	75.70%	76.27%	78.65%

三、狄氏分配之實證研究

（一）Bacteria2035 資料檔之分類正確率比較

圖 1 為 Bacteria2035 使用狄氏分配調整前後之分類正確率比較圖，圖中狄氏分配_分子分母代表的是在各個類別值中，進行調整的順序為分子先驗分配參數先行調整，調整完畢後再調整分母先驗分配參數；而狄氏分配_分母分子，則是先行調整分母的先驗分配參數，再調整分子的先驗分配參數。由此圖可以觀察出狄氏分配_分母分子由 95.57%提升至 98.37%，提升了將近 3 個百分點；而狄氏分配_分子分母相較於狄氏分母分子的分類正確率來的高，由 95.57%提升至 98.88%，提升超過了 3 個百分點，其中藉由先驗分配參數的調整正確率提升較多的原因為狄氏分配_分子分母相較於狄氏分配_分母分子多調整了一個類別值，進而使最後的分類

正確率有較高的結果。

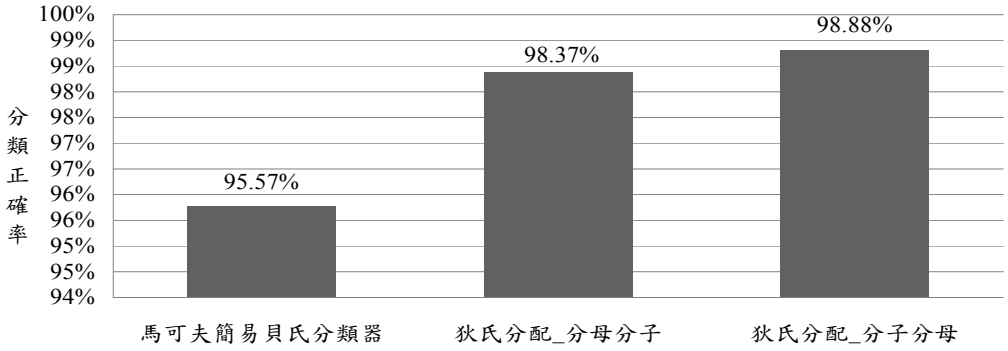

圖 1：Bacteria2035 狄氏分配參數調整前後之正確率比較圖

圖 2 為狄氏分配_分子分母與狄氏分配_分母分子的正確率變化比較圖，分類正確率由初始值 0.000001 時的 95.77%提升至 98.37%和 98.88%，上升約為 3 個百分點。圖 3 與圖 4 則是狄氏分配_分母分子和狄氏分配_分母分子，在各個有效改善之類別下的先驗分配參數，其中狄氏分配_分子分母的 7-mer 分母先驗分配參數在第一個類別值下有較高的調整，並且相較於狄氏分配_分母分子多調整了一個類別值，因此最後會有較佳的分類正確率。而圖中可以觀察出 8-mer 分子的先驗分配參數在類別值編號 12 之前有較多的調整次數，而 7-mer 分母的先驗分配參數則是在從類別值編號 1 開始，都對類別值有陸續的調整。

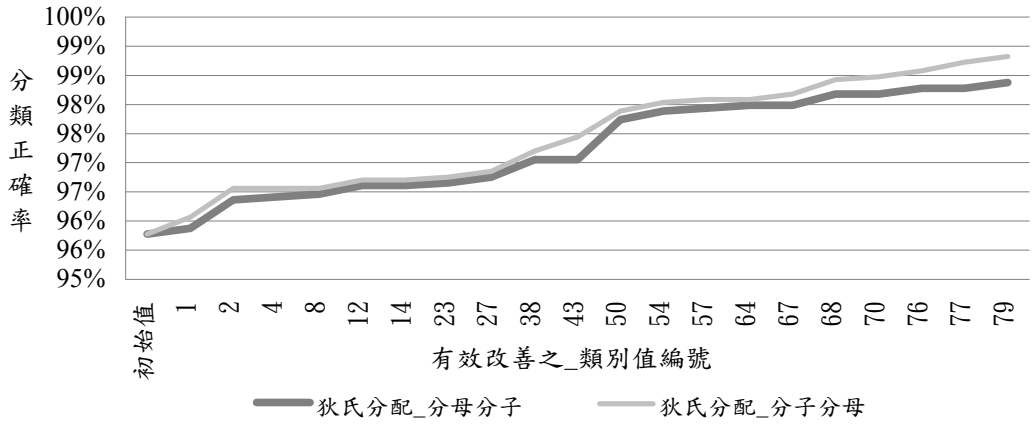

圖 2：Bacteria2035 狄氏分配_分子分母與分母分子之正確率變化比較圖

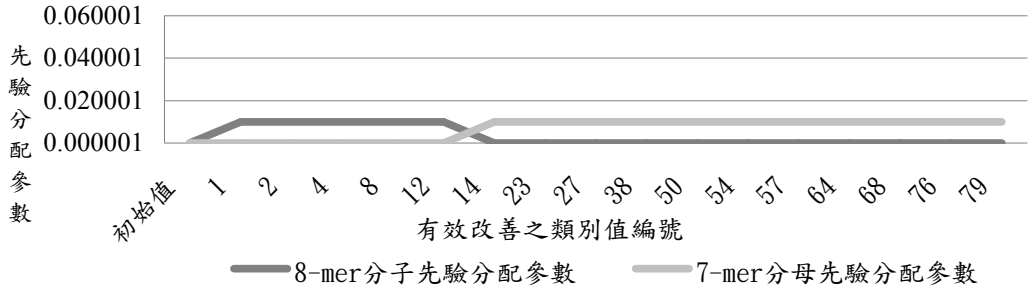

圖 3：Bacteria2035 狄氏分配_分母分子之參數調整變化圖

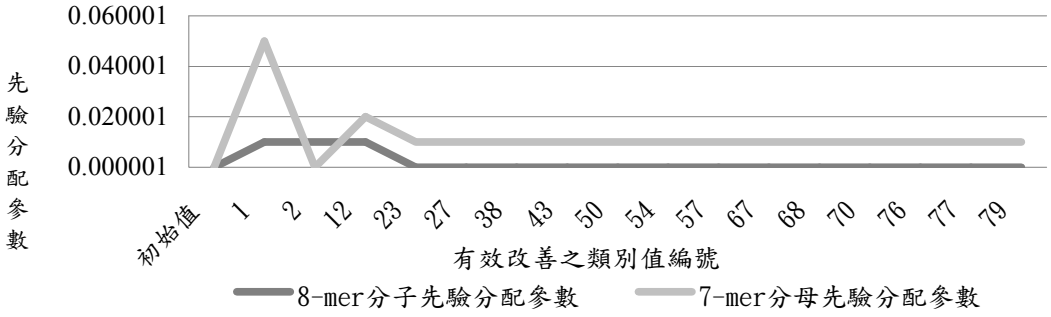

圖 4：Bacteria2035 狄氏分配_分子分母之參數調整變化圖

狄氏分配_分子分母，在類別值獲得有效調整的總數為 17 個，其中 8-mer 分子有效參數調整次數為 9 次，7-mer 有效參數調整次數為 16 次；而狄氏分配_分母分子，在 8-mer 分子有效參數調整次數為 5 次，7-mer 有效參數調整次數為 11 次，類別值調整總數為 17，從中可以發現該兩種方法在分子參數的調整上有較多的調整次數，而分母先驗分配參數的調整次數則是低於分子先驗分配參數的調整次數。

（二）Bacteria3672 資料檔之分類正確率比較

圖 5 為 Bacreria3672 資料檔透過狄氏分配調整參數前後之分類正確率比較圖。在原始參數 0.000001 時，分類正確率為 90.08%，而狄氏分配_分母分子和狄氏分配_分子分母在調整後之分類正確率分別為 96.51%和 97.27%，提升的幅度約為 6 至 7 個百分點，在最終參數的調整下，狄氏分配_分子分母之分類正確率較狄氏分配_分母分子來得高。

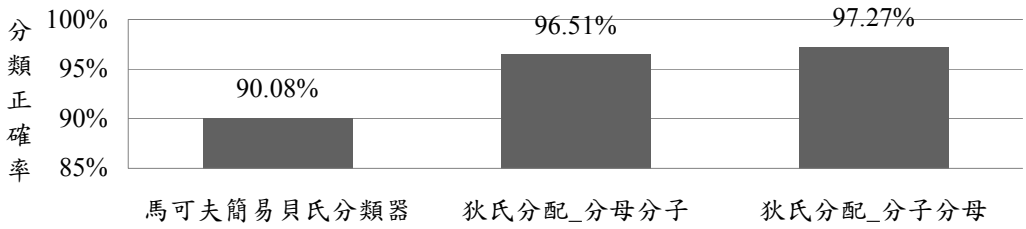

圖 5：Bacteria3672 狄氏分配參數調整前後之正確率比較圖

圖 6 為狄氏分配_分母分子與狄氏分配_分子分母之分類正確率變化比較圖，從圖中可以觀察出該兩種調整方法之分類正確率差異並不大，最大差距為 0.05%。在參數調整過程中，狄氏分配_分子分母從類別值編號 1 的分類正確率就高於狄氏分配_分母分子 0.01%，其中的原因為狄氏分配_分子分母的 7-mer 先驗分配參數在類別值 1 的提升幅度較大，並且在類別值編號 34 之前 7-mer 參數都有陸續的調整，因此狄氏分配_分子分母在往後類別值的調整下都與狄氏分配_分母分子的分類正確率保持一定的差距。

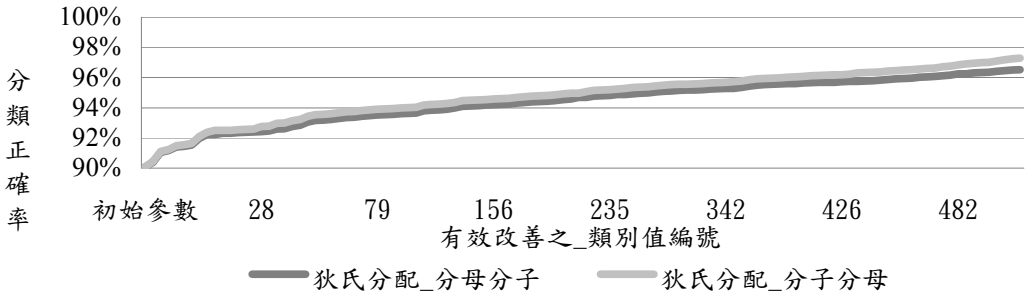

圖 6：Bacteria3672 狄氏分配_分子分母與分母分子之正確率變化比較圖

圖 7 和圖 8 為資料檔 Bacteria3672 在有效改善之類別值編號下的先驗分配參數。狄氏分配_分母分子和狄氏分配_分子分母，在 8-mer 分子先驗分配參數幅度差異不大，而對於 7-mer 分母先驗分配參數而言，則是狄氏分配_分子分母在類別值 1、12、28 和 34 都比狄氏分配_分母分子才有更多的調整幅度，並且比 8-mer 分子先驗分配參數的調整次數多。

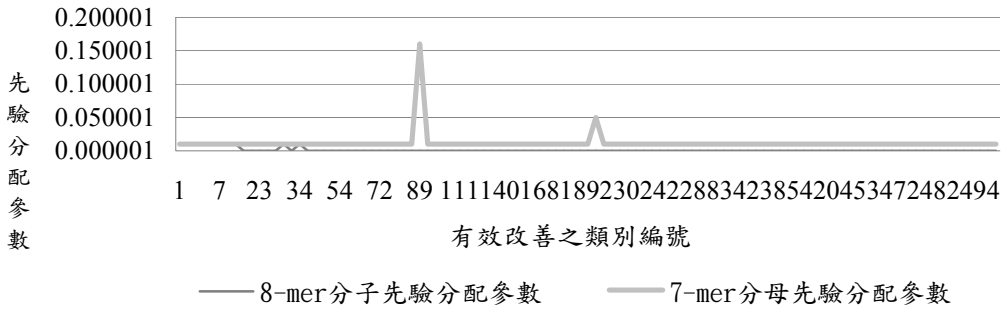

圖 7：Bacteria3672 狄氏分配_分母分子之參數調整變化圖

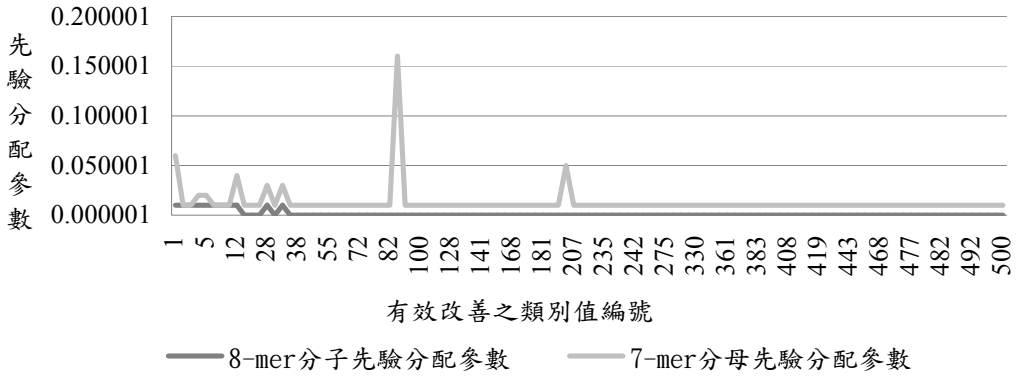

圖 8：Bacteria3672 狄氏分配_分子分母之參數調整變化圖

狄氏分配_分子分母與狄氏分配_分母分子，在參數的調整過程中，狄氏分配_分母分子 8-mer 分子參數的調整次數為 10 次，7-mer 分母調整次數為 103 次，類別值調整總數為 103 次；而狄氏分配_分子分母的 8-mer 分子之調整次數為 11，7-mer 參數的調整次數為 101，類別值調整總數為 109 其中 7-mer 先驗分配參數對於每個類別值都有調整。就調整次數而言，狄氏分配_分子分母在調整次數上雖然比狄氏分配_分母分子少調整一次，但是在類別值的調整個數上與 7-mer 分母參數都有較大的調整，而使最後狄氏分配_分子分母會有較高的分類正確率。

（三）Fungi4954 資料檔之分類正確率比較

圖 9 為 Fungi4954 使用狄氏分配調整前後之正確率比較圖，Fungi4954 在原始初始值參數 0.000001 下之分類正確率為 87.90%，而調整所有參數後，狄氏分配_分母分子和狄氏分配_分子分母之分類正確率分別為 93.78%與 94.57%，正確率上

升了大約 6 個百分點，其中由最後的實驗結果得知狄氏分配_分子分母的分類正確率較高。

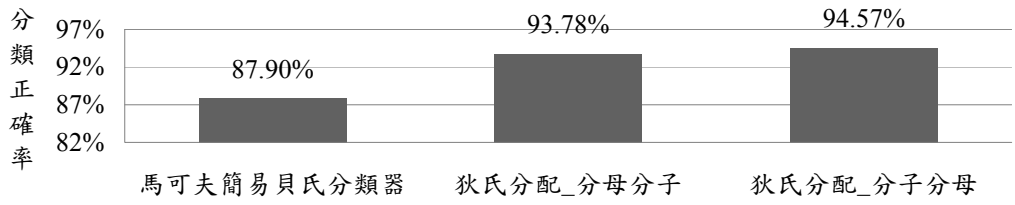

圖 9：Fungi4954 狄氏分配參數調整前後之正確率比較圖

圖 10 為 Fungi4954 狄氏分配_分子分母和狄氏分配_分母分子之間的正確率變化比較圖。狄氏分配_分子分母與狄氏分配_分母分子在類別值 90 之前分類正確率都較為相近，而在類別值編號 90 的參數調整上，狄氏分配_分子分母的分類正確率有較大幅度的提升，而使最終的分類結果狄氏分配_分子分母的分類正確率相對於狄氏分配_分母分子要來得高。

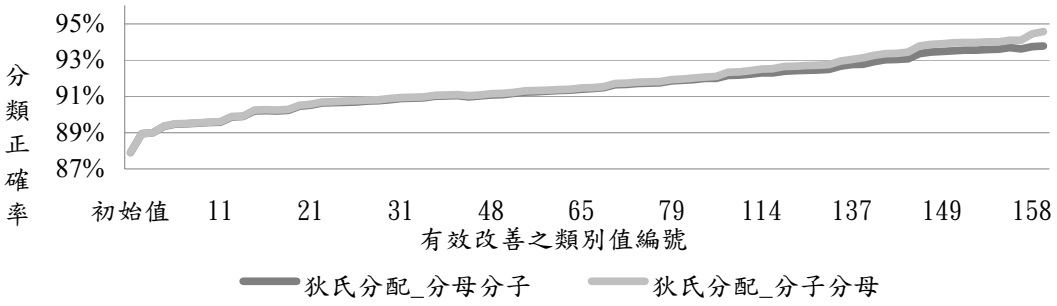

圖 10：Fungi4954 狄氏分配_分子分母與分母分子之正確率變化比較圖

圖 11 和圖 12 則是資料檔 Fungi4954 在各有效改善之類別值下的先驗分配參數，其中狄氏分配_分母分子的 7-mer 分母的先驗分配參數在類別值 2 就開始調整，而狄氏分配_分子分母的 7-mer 分母參數調整頻率較低，8-mer 分子先驗分配參數則是較狄氏分配_分母分子調整幅度來得要高。

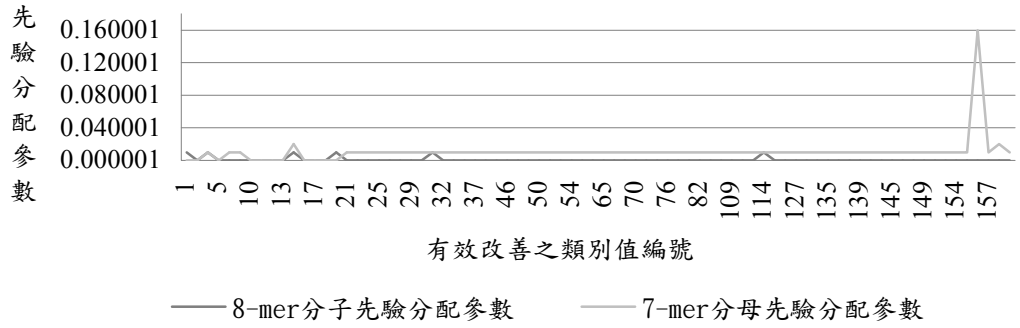

圖 11：Fungi4954 狄氏分配_分母分子之參數調整變化圖

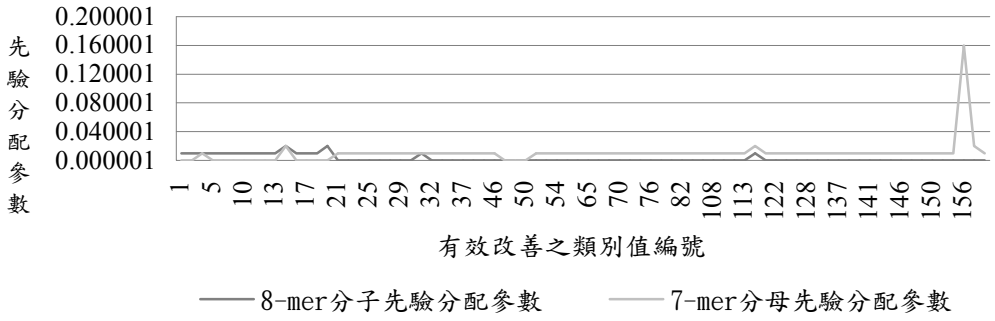

圖 12：Fungi4954 狄氏分配_分子分母之參數調整變化圖

狄氏分配_分子分母與狄氏分配_分母分子，在分子先驗分配與分母先驗分配參數的調整次數與調整總數都不相同。狄氏分配_分母分子的 8-mer 分子調整個數為 7，7-mer 分母調整數目為 67，類別值調整的個數為 78；而狄氏分配_分子分母，在 8-mer 分子調整個數為 17，7-mer 分母調整數目為 65，類別值調整的個數為 78。該兩種調整方法在類別值調整個數上相同，但是在 8-mer 參數的調整上狄氏分配_分子分母有較多的調整，因此會有較佳的分類結果。

（四）Fungi7730 資料檔之分類正確率比較

圖 13 為馬可夫簡易貝氏分類器結合狄氏分配於 Fungi7730 資料檔之參數調整前後的分類正確率比較圖。由該圖中可以發現狄氏分配_分子分母和狄氏分配_分母分子在分類正確率上由原始分類正確率 76.27%上升至 87.30%和 87.94%，提升了將近 11 個百分點，而最終的分類結果為狄氏分配_分母分子有較高的分類正確率。

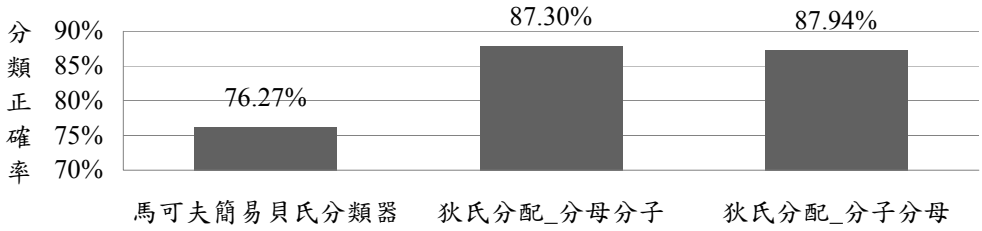

圖 13：Fungi7730 狄氏分配參數調整前後之正確率比較圖

圖 14 為狄氏分配_分子分母與狄氏分配_分母分子的正確率變化比較圖。在先驗分配參數的調整過程中，兩者之分類正確率差異不大，其中狄氏分配_分子分母在類別值編號 8 相對於狄氏分配_分子分母有較大的調整次數，但是在類別值 85 至類別值 101 之間未調整先驗分配參數，使狄氏分配_分母分子於類別值 92 之後超越於狄氏分配_分子分母並維持一定間距直到調整完所有類別值。

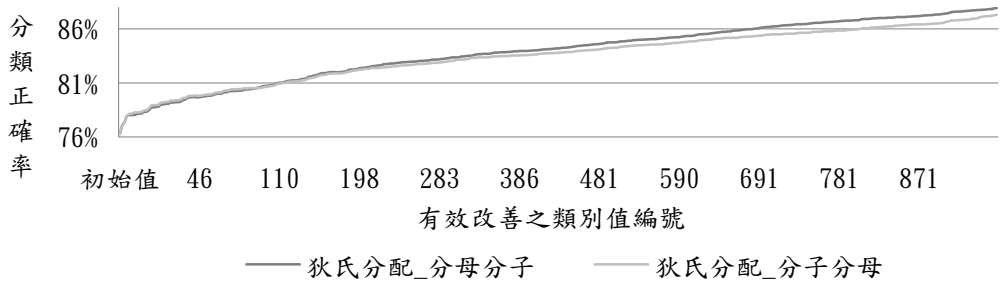

圖 14：Fungi7730 狄氏分配_分子分母與分母分子之正確率變化比較圖

圖 15 和圖 16 為狄氏分配_分母分子和狄氏分配_分子分母在各個有效改善之類別值下的先驗分配參數，其狄氏分配_分母分子的 8-mer 分子先驗分配參數的有效調整次數較狄氏分配_分母分子來得高；而 7-mer 分母先驗分配參數在狄氏分配_分母分子在第一個類別值開始就依續往後調整，並維持參數值 0.010001 直至最後類別值。

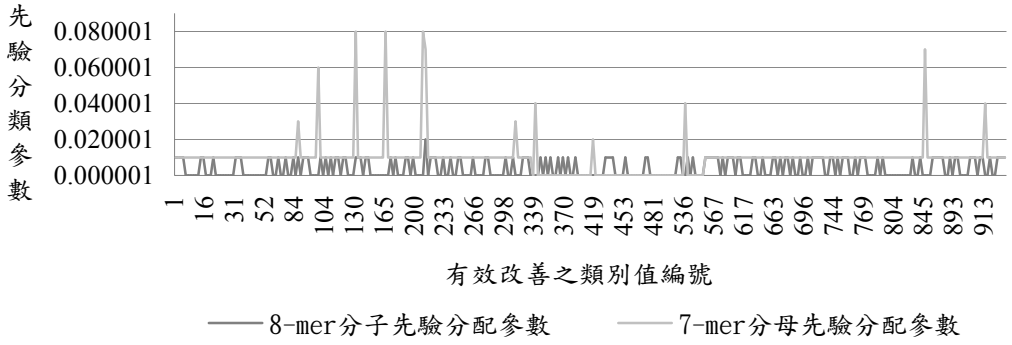

圖 15：Fungi7730 狄氏分配_分母分子之參數調整變化圖

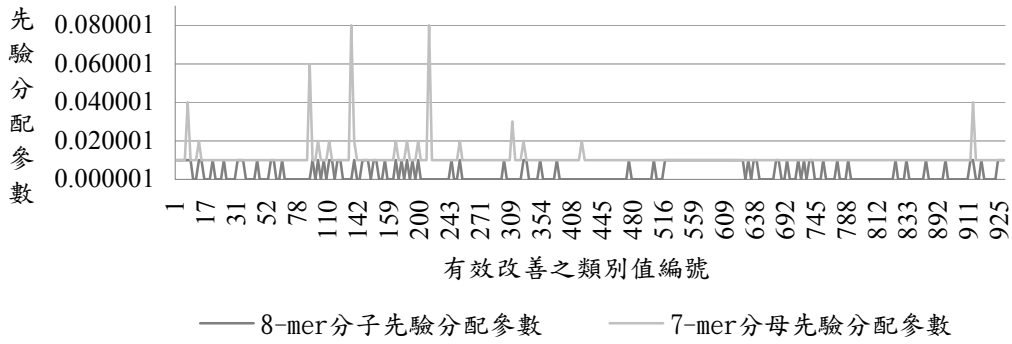

圖 16：Fungi7730 狄氏分配_分子分母之參數調整變化圖

在先驗分配參數的調整過程中，狄氏分配_分母分子的 8-mer 分子先驗分配參數調整次數為 134，7-mer 分母先驗分配參數調整次數為 333，全部類別調整總數為 333，其中狄氏分配_分母分子的 7-mer 分母先驗參數都維持 0.010001；而狄氏分配_分子分母，8-mer 分子先驗分配參數之調整次數為 78，7-mer 分母先驗分配參數調整次數為 299，類別值調整總數為 299。從以上實驗資料可以得知，狄氏分配_分母分子在先驗分配數量與類別值調整的個數比狄氏分配_分子分母要來得多，因此會有較佳的分類正確率。

表 2 和表 3 分別為本研究方法與 RDP 分類器在 Bacteria 和 Fungi 資料檔中各個階層之分類正確率比較。由表可知在「門」和「綱」階層為 RDP 分類器有較佳的分類結果，在「目」階層中只有一個資料檔為 RDP 分類器的分類正確率勝出；然而，在「科」和「屬」這兩個階層，無論是哪個資料檔都為馬可夫簡易貝氏分類器結合狄氏分配有較高的分類正確率，而「目」階層中有三個資料檔也為狄氏

分配_分子分母有較佳的分類正確率。

表 2：各方法於 Bacteria 資料檔下之分類正確率比較

階層	Bacteria 2035				Bacteria 3672			
	RDP 分類器	馬可夫簡易貝氏分類器	狄氏分配_分子分母	狄氏分配_分母分子	RDP 分類器	馬可夫簡易貝氏分類器	狄氏分配_分子分母	狄氏分配_分母分子
界	1	1	1	1	1	1	1	1
門	**1**	0.99459	0.99705	0.99803	**0.99918**	0.98366	0.99482	0.99564
綱	**0.99803**	0.98673	0.99410	0.99410	**0.99265**	0.97630	0.99019	0.98992
目	0.98918	0.98132	**0.99115**	0.99115	0.98420	0.96759	**0.98692**	0.98529
科	0.98869	0.98034	**0.99115**	0.98722	0.98012	0.96241	**0.98474**	0.97984
屬	0.97297	0.95773	**0.98820**	0.98378	0.93464	0.90087	**0.97276**	0.96514

表 3：各方法於 Fungi 資料檔下之分類正確率比較

階層	Fungi 4954				Fungi 7730			
	RDP 分類器	馬可夫簡易貝氏分類器	狄氏分配_分子分母	狄氏分配_分母分子	RDP 分類器	馬可夫簡易貝氏分類器	狄氏分配_分子分母	狄氏分配_分母分子
界	1	1	1	1	1	1	1	1
門	**0.99979**	0.98990	0.99515	0.99414	**0.99961**	0.96714	0.97774	0.98292
綱	**0.99900**	0.98788	0.99333	0.99051	**0.99379**	0.93505	0.96222	0.97244
目	0.98547	0.97597	**0.98788**	0.97900	**0.96145**	0.90996	0.93880	0.94980
科	0.95902	0.95781	**0.98122**	0.97254	0.89935	0.87309	0.91966	**0.92263**
屬	0.88030	0.87908	**0.94570**	0.93782	0.78396	0.76274	0.87309	**0.87943**

表 4 為吳沐穎（2012）所提之簡易貝氏分類器結合狄氏分配方法與本研究方法狄氏分配_分母分子和狄氏分配_分子分母於 Fungi7730 資料檔下的分類正確率比較。表 5 的正確率比較中，可以發現馬可夫簡易貝氏分類器結合狄氏分配，相對於簡易貝氏分類器結合狄氏分配，多了分母先驗分配參數可供調整，而使參數的調整更具有彈性，因此會有較佳的分類正確率。

表 5 為本研究各個實驗方法與 RDP 分類器在各資料檔下的分類正確率比較，其中加入了該資料檔下執行方法時所需的運算時間。RDP 分類器和馬可夫簡易貝氏分類器，在執行時所需的運算時間皆為數個小時即可執行完畢。在馬可夫簡易

貝氏分類器結合狄氏分配後，由於尋找參數的流程次數為類別值個數的兩倍，因此運算的時間複雜度也隨著資料檔中類別值個數的增加而跟著增加。然而加入了狄氏分配後，雖然會使執行時間增加，但是藉由先驗分配參數的設定，使概似機率在加入了先驗分配參數後，能夠推估得更為精準。

表 4：本研究方法與吳沐穎研究方法之分類正確率比較

實驗方法 資料檔	簡易貝氏分類器 結合狄氏分配	狄氏分配_分母分子	狄氏分配_分子分母
Fungi 7730	83.97%	87.94%	87.30%

表 5：本研究方法與 RDP 分類器之分類正確率與時間比較

實驗方法	RDP 分類器		馬可夫簡易貝氏分類器_ 參數 0.000001		狄氏分配_ 分母分子		狄氏分配_ 分子分母	
資料檔	正確率	時間	正確率	時間	正確率	時間	正確率	時間
Bacteria 2035	97.29%	10（時）	95.77%	56（分）	98.37%	5（天）	**98.88%**	5（天）
Bacteria 3672	93.65%	11（時）	90.08%	649（分）	96.51%	27（天）	**97.27%**	27（天）
Fungi 4954	88.02%	10（時）	87.90%	196（分）	93.78%	9（天）	**94.57%**	9（天）
Fungi 7730	78.65%	12（時）	76.27%	1031（分）	**87.94%**	58（天）	87.30%	58（天）

表 6 是四種分類模型在 16S rRNA 細菌資料集測試下所取得正確率的統計檢定結果，其中右上三角是 Bacteria2035 資料集，左下三角則是 Bacteria3672 資料集；表 7 則是四種分類模型在 28S rRNA 真菌資料集的正確率檢定結果，右上三角和左下三角分別是 Fungi4954 與 Fungi7730 資料集的檢定結果，括號數字代表檢定結果為負值，無括號數字則為正值；顯示粗體表示該組合之正確率在 95%信心水準下其檢定結果顯示有顯著差異。從兩個資料集的檢定結果可以發現，本研究所提出的分類模型（狄氏分配_分母分子、狄氏分配_分子分母）在分類正確率都有顯著高於 RDP 分類器以及參數設定為 0.000001 的馬可夫簡易貝式分類器，檢定 Z 值都大於 2，然本研究所提出模型的兩種參數設定方式所得到的分類正確率則沒有顯著的差異（Z=1.4）。

表 6：正確率統計檢定－細菌 16S rRNA 資料集

	RDP 分類器	馬可夫簡易貝氏分類器_參數 0.000001	狄氏分配_分母分子	狄氏分配_分子分母
RDP 分類器	X	**2.66**	**（2.36）**	**（7.24）**
馬可夫簡易貝氏分類器_參數 0.000001	**5.60**	X	**（4.92）**	**（6.15）**
狄氏分配_分母分子	**（5.67）**	**（8.20）**	X	（1.40）
狄氏分配_分子分母	**（7.45）**	**（9.42）**	（1.40）	X

表 7：正確率統計檢定－真菌 28S rRNA 資料集

	RDP 分類器	馬可夫簡易貝氏分類器_參數 0.000001	狄氏分配_分母分子	狄氏分配_分子分母
RDP 分類器	X	0.18	**（9.97）**	**（11.56）**
馬可夫簡易貝氏分類器_參數 0.000001	**3.54**	X	**（10.15）**	**（11.74）**
狄氏分配_分母分子	**（15.48）**	**（18.93）**	X	（1.68）
狄氏分配_分子分母	**（14.31）**	**（17.77）**	1.21	X

伍、結論與未來研究

本研究方法是以馬可夫模型與簡易貝氏分類器作結合，針對基因序列資料進行分類，分類器的機率模型是採用多項式模型，計算出特徵的出現頻率，並且加入了先驗分配－狄氏分配，透過尋找適合的先驗分配參數，期望能夠有效的提升分類正確率。

實證結果發現，馬可夫簡易貝氏分類器加入了 Laplace Estimate 後，在參數值 0.000001 的分類正確率會比參數值 0.01 的分類正確率來得高，原因為參數值設定為 0.01 時，對於眾多可能特徵集合數量 4^N 的基因序列資料而言，信心水準會過高，進而影響分類正確率。馬可夫簡易貝氏分類器在尚未加入狄氏分配前，其分類正確率是低於 RDP 分類器，我們探討可能的原因為馬可夫簡易貝氏分類器相對於簡易貝氏分類器在式子的計算上多了(N-1)-mer 分母概似機率的計算，進而影響原先 N-mer 分子機率值的計算。在馬可夫模型結合狄氏分配後，其缺點為執行時間會因

為資料檔的類別值個數不同，而使運算時間隨著類別值個數的增加而隨著增加。因此若是以分類正確率作為優先考量時，則可以採用本論文方法；相反地，若是以時間作為優先考量，則能夠採用 RDP 分類器。然而透過先驗分配參數的調整，在四個資料檔下其分類正確率已高於 RDP 分類器，其中在 Fungi7730 的資料檔中，相較於吳沐穎（2012）所提之方法會有較好的分類表現，其原因為本研究採用馬可夫簡易貝氏分類器，相對於簡易貝氏分類器結合狄氏分配多了分母先驗分配參數可供調整，因此可進行調整的參數個數在每個類別值中會增加 N-2 個，使最後的驗證結果會有較好表現。在分子先驗分配參數與分母先驗分配參數的調整順序上，本研究以兩種不同的方式－狄氏分配_分子分母和狄氏分配_分母分子進行測試，實證結果發現狄氏分配_分子分母，在同一個類別值內先進行分子參數的調整，再進行分母參數的調整會有較好的分類結果。故本研究方法多項式馬可夫簡易貝氏分類器結合狄氏分配，在使用相同的資料檔下，尋找出適合的先驗分配參數，確實對分類正確率能夠有效的提升，以上即為本研究的貢獻所在。

由於本研究是對同一個類別內兩種先驗分配參數進行不同順序的調整，未來可以嘗試對整個資料檔的所有類別值先進行某一種先驗分配參數的調整，在調整完畢後再進行另一種先驗分配參數的調整，以找出最好的調整方法。

本研究是使用馬可夫簡易貝氏分類器結合狄氏分配，在未來也許可以採用廣義狄氏分配對同一個類別內不同的變數進行不同參數的調整，使同一個類別值內的變數擁有不同的信心水準，讓調整參數時更具有彈性，以取得更好的分類正確率。

誌謝

本研究之經費由國科會編號 NSC 102-2410-H-006-076-MY2 之計畫所贊助。

參考文獻

吳沐穎(2012)，『簡易貝氏分類器中廣義狄氏先驗分配應用於基因序列資料分類之研究』，未出版碩士論文，國立成功大學資訊管理研究所，台南市。

Baum, L.E. and Eagon, J.A. (1967), 'An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology', *Bulletinof theAmerican Mathematical Society*, Vol. 73, No. 3, pp. 360-363.

Brady, A. and Salzberg, S.L. (2009), 'Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models', *Nature Methods*, Vol. 6, No. 9, pp. 673-678.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009), 'BLAST+: architecture and applications', *BMC Bioinformatics*, Vol. 10, No. 1, pp. 421.

Chen, J., Huang, H., Tian, S. and Qu, Y. (2009), 'Feature selection for text classification with Naïve Bayes', *Expert Systems with Applications*, Vol. 36, No. 3, pp. 5432-5435.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R. and Tiedje, J.M. (2014), 'Ribosomal database project: Data and tools for high throughput rRNA analysis', *Nucleic Acids Research*, Vol. 42, No. 1, pp. 633-642.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K. and Andersen, G.L. (2006), 'Greengenes, a chimera-checked 16S rRNAgene database and workbench compatible with ARB', *Applied and Environmental Microbiology*, Vol. 72, No. 7, pp. 5069-5072.

Ghosh, T., Gajjalla, P., Mohammed, M. and Mande, S. (2012), 'C16S - A hidden Markov model based algorithm for taxonomic classification of 16S rRNA gene sequences', *Genomics*, Vol. 99, No. 4, pp. 195-201.

Gilbert, J.A. and Meyer, F. (2012), 'Modeling the Earth's microbiome: a real world deliverable for microbial ecology', *ASM Microbe Magazine*, Vol. 7, No. 2, pp. 64-69.

Handelsman, J., Rondon, M.R., Brady, S., Clardy, J. and Goodman, R.M. (1998), 'Molecular biology provides access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & Biology*, Vol. 5, No. 10, pp. 245-249.

Human Microbiome Project Consortium (2012), 'A framework for human microbiome research', *Nature*, Vol. 486, No. 7402, pp. 215-221.

Kotamarti, R., Hahsler, M. and Raiford, D. (2010), 'Analyzing taxonomic classification using extensible Markov models', *Bioinformatics*, Vol. 26, No. 18, pp. 2235-2241.

Liu, K.L. and Wong, T.T. (2013), 'Naïve Bayesian classifiers with multinomial models for rRNA taxonomic assignment', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 10, No. 5, pp. 1334-1339.

Liu, K.L., Porras-Alfaro, A., Kuske, C.R., Eichorst, S.A. and Xie, G. (2012), 'Accurate, rapid taxonomic classification of fungal Large-Subunit rRNA genes', *Applied and Environmental Microbiology*, Vol. 78, No. 5, pp. 1523-1533.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A. and Schleifer,

K.H. (2004), 'ARB: a software environment for sequence data', *Nucleic Acids Research*, Vol. 32, No. 4, pp. 1363-1371.

McCallum, A. and Nigam, K. (1998), 'A comparison of event models for naive Bayes text classification', *Working Notes of the 1998 AAAI/ICML Workshop on Learning for Text Categorization*, Vol. 752, pp. 41-48.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R. and Stahl, D.A. (1986), 'Microbial ecology and evolution - a ribosomal-RNA approach', *Annual Review of Microbiology*, Vol. 40, No. 1, pp. 337-365.

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. and Sokhansanj, B. (2008), 'Metagenome fragment classification using N-mer frequency profiles', *Advances in Bioinformatics*, Vol. 2008, pp. 1-12.

Rosen, G.L., Reichenberger, E.R. and Rosenfeld, A.M. (2011), 'NBC: the naive Bayes classification tool webserver for taxonomic classification of metagenomic reads', *Bioinformatics*, Vol. 27, No. 1, pp. 127-129.

Wang, Q., Garrity, G.M., Tiedje, J.M. and Cole, J.R. (2007), 'Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy', *Applied and environmental microbiology*, Vol. 73, No. 16, pp. 5261-5267.

Woese, C.R. (1987), 'Bacterial evolution', *Microbiological Reviews*, Vol. 51, No. 2, pp. 221-271.

Woese, C.R. and Fox, G.E. (1977), 'Phylogenetic structure of prokaryotic domain-primary Kingdoms', *Proceedings of the National Academy of Sciencesof the United States of America*, November 1, Vol. 74, No. 11, pp. 5088-5090.

Wong, T.T. (1998), 'Generalized Dirichlet distribution in Bayesian analysis', *Applied Mathematics and Computation*, Vol. 97, No. 2, pp. 165-181.

Wong, T.T. (2009), 'Alternative prior assumptions for improving the performance of naïve Bayesian classifiers', *Data Mining and Knowledge Discovery*, Vol.18, No. 2, pp. 183-213.